The Annals of Statistics
2014, Vol. 42, No. 3, 1102–1130
DOI: 10.1214/14-AOS1215

POSTERIOR CONTRACTION IN SPARSE BAYESIAN FACTOR MODELS FOR MASSIVE COVARIANCE MATRICES

By Debdeep Pati[1], Anirban Bhattacharya[1], Natesh S. Pillai[2] and David Dunson[1]

Florida State University, Texas A&M University, Harvard University and Duke University

Sparse Bayesian factor models are routinely implemented for parsimonious dependence modeling and dimensionality reduction in high-dimensional applications. We provide theoretical understanding of such Bayesian procedures in terms of posterior convergence rates in inferring high-dimensional covariance matrices where the dimension can be larger than the sample size. Under relevant sparsity assumptions on the true covariance matrix, we show that commonly-used point mass mixture priors on the factor loadings lead to consistent estimation in the operator norm even when $p \gg n$. One of our major contributions is to develop a new class of continuous shrinkage priors and provide insights into their concentration around sparse vectors. Using such priors for the factor loadings, we obtain similar rate of convergence as obtained with point mass mixture priors. To obtain the convergence rates, we construct test functions to separate points in the space of high-dimensional covariance matrices using insights from random matrix theory; the tools developed may be of independent interest. We also derive minimax rates and show that the Bayesian posterior rates of convergence coincide with the minimax rates upto a $\sqrt{\log n}$ term.

1. Introduction. It is now routine to collect data where the dimension p is much larger than the sample size n, and interest focuses on the covariance structure. In this context, even a simple parametric model like the

Received September 2013; revised February 2014.
[1]Supported in part by the DARPA MSEE program and Grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).
[2]Supported in part by NSF Grant DMS-11-07070.
AMS 2000 subject classifications. 62G05, 62G20.
Key words and phrases. Bayesian estimation, covariance matrix, factor model, rate of convergence, shrinkage, sparsity.

Gaussian distribution leads to a high-dimensional model space and it becomes necessary to reduce the effective number of parameters via imposing sparsity or some lower-dimensional structure. Sparse Bayesian factor models [41] provide one popular choice in applications, but currently lack theoretical support. In this paper, we close this gap by studying asymptotic properties for scenarios in which p grows faster than n.

Factor models [5] aim to explain dependence among multivariate observations through shared dependence on a smaller number of latent factors. Given n i.i.d. observations $y_i \in \mathbb{R}^p$, a latent factor model is given by

$$y_i = \Lambda \eta_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}_p(0, \Omega), i = 1, \ldots, n, \tag{1.1}$$

where Λ is a $p \times k$ factor loadings matrix with $k \ll p$, $\eta_i \sim \mathrm{N}_k(0, \mathrm{I})$ are standard normal latent factors, and ε_i is a residual having diagonal covariance $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Marginalizing out the latent factors, $y_i \sim \mathrm{N}_p(0, \Sigma)$ with

$$\Sigma = \Lambda\Lambda^{\mathrm{T}} + \Omega, \tag{1.2}$$

so that the right-hand side has at most $p(k+1)$ parameters compared to $O(p^2)$ parameters in an unstructured covariance matrix.

A prior distribution on (Λ, Ω) induces a prior distribution on Σ and we are interested in studying concentration of the corresponding posterior measure around a "true" covariance matrix in *operator norm* when the dimensionality $p = p_n$ can be much larger than the sample size n. This setting has motivated abundant frequentist work, with rates of convergence of various regularized covariance estimators derived in [7, 8, 11, 12, 18, 30] among others. Minimax optimal rates for specific sparsity classes have also been derived in [13, 14]. There is a relatively smaller but increasing literature on asymptotic properties of Bayesian procedures in models with growing dimension, primarily focused on linear or generalized linear models; refer to [2, 6, 10, 17, 22, 23] among others. To the best of our knowledge, the present paper is the first to study the asymptotic properties of Bayesian covariance estimation via factor models in the $p_n \gg n$ regime.

We now summarize the main results obtained in this paper. Although the original specification of the factor model reduces the number of parameters from quadratic to linear in p_n, the estimation problem is still challenging when $p_n \gg n$. To address this challenge, [41] introduced *sparse factor modeling* to allow many of the loadings to be exactly equal to zero through a point mass mixture prior having a probability mass at zero; see also [15, 33] for modifications and applications in genomics. Recently, [17] studied posterior concentration in estimating a sparse high-dimensional mean using such point mass mixture priors. However, it is not clear whether the *induced* prior on the covariance from such sparsity favoring priors on the factor loadings would lead to consistent covariance estimation in the $p_n \gg n$ setting. We answer the question in the affirmative and derive the rate of convergence of the posterior in Section 5, explicitly characterizing the dependence on the dimensionality p_n, the true number of factor k_{0n}, the column sparsity s_n in the true loadings and the growth rate of the largest eigenvalue c_n of the true covariance. In particular, the dimensionality enters the rate through a logarithmic factor, providing justification of usage of such methods in *ultra high-dimensional* settings. It may be remarked here that the usual practice of assuming the eigenvalues of the true covariance to be bounded is restrictive in our context and we relax that assumption.

Although point mass mixture priors are amenable to incorporate sparsity, exploring the model space via MCMC can be daunting and may lead to slow mixing and convergence of the algorithm [36]. To address such problems through block updating, while allowing a weaker notion of sparsity in which elements are close to zero instead of exactly zero, continuous shrinkage priors can be used. Such priors have become common in regression [2, 16, 28, 35], with [36] providing a unifying local–global scale mixture representation. Although computationally attractive, the lack of tight concentration bounds for such priors has limited the study of their asymptotic properties. One of our main contributions is to develop a novel class of continuous shrinkage priors and derive nonasymptotic bounds on the concentration and dimensionality of such priors. Based on these results, we show that the proposed continuous shrinkage prior leads to the same rate of posterior convergence as the point mass mixture priors in estimating large covariance matrices.

The Birgé–Le Cam testing theory [9, 32] for the Hellinger metric is commonly used in Bayesian asymptotics [24] to separate points in the parameter space. However, generalization of the testing argument to other norms has been relatively unexplored. A notable exception is [26] who advocated the use of concentration inequalities based on empirical process techniques to derive tests in the L_r metric in a nonparametric function estimation context. See also [37] for an usage of concentration bounds for centered linear estimators in the context of test construction in Bayesian inverse problems. In the setting of large covariance estimation in *operator norm*, we construct tests inspired by results from the nonasymptotic theory of random matrices, which might be of independent interest in related settings.

Finally, we use Fano's lemma to derive the *minimax rate* of convergence for the class of covariance matrices considered in this paper and show that the posterior indeed convergences at the minimax rate up to a $\sqrt{\log n}$ term.

There is a sizeable literature studying asymptotic properties of various aspects of factor analysis, including consistent estimation of factor loadings and latent factors [3] and the number of factors [4, 31]. Fan, Fan and Lv [19] studied rates of convergence of high-dimensional covariance estimates based on factor models, with [20] extending their results to approximate factor models that allow nondiagonal Ω in (1.2). This work assumes that the factor scores η_i are known, while we consider the fundamentally different setting in which the factor scores are unknown while also studying concentration of a Bayesian posterior instead of convergence of a point estimate.

The rest of the paper is organized as follows. After setting up the basic notation and definitions in Section 2, we state our assumptions and their implications in Section 3. In Section 4, we discuss our prior distributions. The main results of this paper are stated in Section 5. Section 6 contains some numerical simulations. In Section 7, we prove a number of concentration bounds for the shrinkage prior introduced in Section 4, while in Section 8, we elucidate our test construction. These results are used to prove the main results in Section 9. Proof of some technical lemmas are given in a supporting document.

2. Preliminaries. Given sequences a_n, b_n, we shall denote $a_n = O(b_n)$ or $a_n \lesssim b_n$ if there exists a global constant C such that $a_n \leq C b_n$. Similarly, we define $a_n \gtrsim b_n$ and $a_n \asymp b_n$.

Given a metric space (X, d), let $N(\varepsilon; X, d)$ denote its ε-covering number, that is, the minimum number of balls of radius ε needed to cover X.

For a vector $x \in \mathbb{R}^r$, $\|x\|_2$ denotes its Euclidean norm. We will use $\mathcal{S}^{r-1}$ to denote the unit Euclidean sphere $\{x \in \mathbb{R}^r : \|x\|_2 = 1\}$ and Δ^{r-1} to denote the $(r-1)$-dimensional simplex $\{x = (x_1, \ldots, x_r)^{\mathrm{T}} : x_j \geq 0, \sum_{j=1}^r x_j = 1\}$. Further, let Δ_0^{r-1} denote $\{x = (x_1, \ldots, x_{r-1})^{\mathrm{T}} : x_j \geq 0, \sum_{j=1}^{r-1} x_j \leq 1\}$.

For a square matrix A, $\operatorname{tr}(A)$ and $|A|$, respectively, denote the trace and the determinant of A. For a $p \times r$ matrix $A = (a_{jj'})$ with $p \geq r$, let $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(r)} \geq 0$ denote the singular values of A (or equivalently the eigenvalues of $\sqrt{A^{\mathrm{T}} A}$) arranged in decreasing order. We shall use $s_{\min}(A)$ and $s_{\max}(A)$ to denote the smallest and largest singular values, respectively. The Frobenius norm ($\|\cdot\|_F$) and the operator norm ($\|\cdot\|_2$) are defined in the usual way, with $\|A\|_F := \sqrt{\operatorname{tr}(A^{\mathrm{T}} A)}$ and $\|A\|_2 := \sup_{x \in \mathcal{S}^{r-1}} \|Ax\|_2 = s_{\max}(A)$. Also $\|A\|_1 = \sum_{j=1}^p \sum_{h=1}^r |A_{jh}|$ is the l_1 norm of $\operatorname{vec}(A)$. We will derive posterior convergence rates in the *operator norm*.

For a subset $S \subset \{1, \ldots, p\}$, let $|S|$ denote the cardinality of S and define $\theta_S = (\theta_j : j \in S)$ for a vector $\theta \in \mathbb{R}^p$. Denote $\operatorname{supp}(\theta)$ to be the *support* of θ, that is, the subset $S_0 \subset \{1, \ldots, p\}$ corresponding to the nonzero entries of θ. We shall continue to use the same notation for a subset of entries and support for matrices Λ, where it has to be interpreted that Λ is vectorized column-wise. Let $l_0[s; p]$ be the space of s-sparse vectors $\theta \in \mathbb{R}^p$ with $|\operatorname{supp}(\theta)| \leq s$.

Throughout C, C' are generically used to denote positive constants whose values might change from one line to the next but are independent from everything else.

Finally, let $\mathcal{C}_n$ denote the cone of covariance matrices of size $p_n \times p_n$ and let $\Sigma_{0n} \in \mathcal{C}_n$ denote a true sequence of covariance matrices.[3] We observe

$$y_1, \dots, y_n \overset{\text{i.i.d}}{\sim} \mathrm{N}_{p_n}(0, \Sigma_{0n})$$

and set $\mathbf{y}^{(n)} = (y_1, \dots, y_n)$. We model the data as

$$y_i \overset{\text{i.i.d}}{\sim} \mathrm{N}_{p_n}(0, \Sigma_n), \qquad \Sigma_n = \Lambda_n \Lambda_n^{\mathrm{T}} + \Omega_n, \Omega_n = \sigma^2 \mathrm{I}_{p_n}. \tag{2.1}$$

We will denote our prior distribution on $\mathcal{C}_n$ (constructed in Section 4) by $\Pi_n(\cdot)$ and the corresponding posterior distribution by $\Pi_n(\cdot|\mathbf{y}^{(n)})$.

3. Assumptions. In this section, we state our assumptions on the true data generating model and briefly discuss their implications. Let $\mathbb{R}^{p\times k}$ denote the class of real-valued $p \times k$ matrices. We start with the following assumptions on the true covariance matrix of the observed data $\mathbf{y}^{(n)}$.

Assumption 3.1. The true sequence of covariance matrices Σ_{0n} are of the form

(A0) $\Sigma_{0n} = \Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} + \Omega_{0n}, \qquad \Lambda_{0n} \in \mathbb{R}^{p_n \times k_{0n}}, k_{0n} \le p_n, \Omega_{0n} = \sigma_{0n}^2 \mathrm{I}_{p_n}.$

Assumption (3.1) says that the true sequence of covariances Σ_{0n} admit a factor decomposition as in (1.2) with $\Omega_{0n} = \sigma_{0n}^2 \mathrm{I}_{p_n}$. We make the following assumptions on Λ_{0n} and σ_{0n}^2.

Assumption 3.2. There exist sequences of positive real numbers c_n, s_n with $c_n \lesssim s_n$, such that:

(A1) $\lim_{n\to\infty} c_n k_{0n}^{3/2} \sqrt{\frac{s_n \log p_n}{n}} \sqrt{\log n} = 0$; $k_{0n}^{3/2} \sqrt{\frac{s_n \log p_n}{n}} (\log n)^{3/2} = O(1)$.
(A2) Each column of Λ_{0n} belongs to $l_0[s_n; p_n]$.
(A3) $\|\frac{1}{c_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n} - \mathrm{I}_{k_{0n}}\|_2 = o(k_{0n}\sqrt{\log k_{0n}/n})$.
(A4) There exists a constant $\sigma_0^{(1)}$ such that $\sigma_0^{(1)} \le \sigma_{0n}^2 \le c_n$.

We now discuss implications of each of the above assumptions.

- If $k_{0n} = O(1)$, $c_n, s_n \asymp \log p_n$, the first part of (A1) allows p_n to grow faster than n under the mild assumption of $(\log p_n)^5 \log n/n \to 0$. In this case, p_n can be of the order of $\exp(n^\alpha)$ for any $\alpha \in (0, 1/5)$. The second part is a very mild requirement given the first part; indeed if $p_n = \exp(n^\alpha)$, $s_n = n^\beta$ and $k_{0n} = n^\gamma$ for appropriate $\alpha, \beta, \gamma > 0$ such that the first part of (A1) holds, then the second part follows from the first.

[3] As a convention, we make the dependence of all quantities on n explicit, and only omit that in a few places for notational convenience.

- In gene-expression studies, we expect each factor is related to only a relatively small number of variables, representing a sparse, parsimonious structure underlying the associations among genes. Following the motivation in [41], usually a small number of latent factors associate with the response so that only those genes with nonzero loadings on those factors are relevant. This is reflected through (A2), requiring the loadings columns to be sparse with $s_n \ll p_n$ many signals per column.
- Conditions similar to (A3) appear in the econometric factor model setting [19, 20] referred to as "pervasive." We provide an intuition based on random matrix theory which suggests that (A3) is indeed mild and expected to be satisfied by a large class of loadings. As our emphasis is on sparse factor models, a realistic generative model for the true loadings would be

$$\lambda_{0jh} \sim (1-\pi_n)\delta_0 + \pi_n \mathrm{N}(0,1),$$

where $\lambda_{0jh} = [\Lambda_{0n}]_{jh}$, δ_0 denotes a point mass at zero and we set $\pi_n = s_n/p_n$ to reflect the sparsity assumption in (A2). Using a modification of Theorem 5.39 of [40], $\|\frac{1}{p_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n} - \pi_n \mathrm{I}_{k_{0n}}\|_2 \leq C\frac{\sqrt{k_{0n}}}{\sqrt{p_n}}\|\pi_n \mathrm{I}_{k_{0n}}\|_2$, or equivalently, $\|\frac{1}{s_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n} - \mathrm{I}_{k_{0n}}\|_2 \leq C\frac{\sqrt{k_{0n}}}{\sqrt{p_n}}$, with probability at least $1 - e^{-C' k_{0n}}$. We can thus choose $c_n = s_n$ and $\sqrt{k_{0n}/p_n}$ is smaller than $k_{0n}\sqrt{\log k_{0n}/n}$ if $p_n = \exp(n^\alpha)$.
- (A4) simply posits an upper and lower bound on the residual variance. The lower bound is used to avoid Σ_{0n} being ill-conditioned,[4] while the upper bound ensures that the larger contribution to $\|\Sigma_{0n}\|_2$ comes from the loadings Λ_{0n}. In particular, (A3) and (A4) imply $\|\Sigma_{0n}\|_2 \asymp c_n$, allowing the largest eigenvalue to grow with increasing dimension.

We denote by $\mathcal{C}_{0n}$ the class of covariance matrices satisfying (A0)–(A4) in Assumptions 3.1 and 3.2. Clearly, any $\Sigma_{0n} \in \mathcal{C}_{0n}$ can be parameterized by $(k_{0n}, \Lambda_{0n}, \sigma_{0n}^2)$, where $\Lambda_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$.

4. Prior distribution. We consider model (2.1) with $\Lambda_n \in \mathbb{R}^{p_n \times k}$. We specify priors on the residual variance σ^2, the number of factors k and the factor loadings (conditional on the number of factors) $\Lambda_n \mid k$ below.

For the residual variance σ^2, we assign a gamma prior f_σ on $(0,\infty)$,

$$\sigma^2 \sim \mathrm{Ga}(a,b). \tag{PR}$$

For the number of factors k, we assume a prior distribution π_k which decays exponentially,

$$\pi_k(k > j) \leq \exp(-Cj), \tag{4.1}$$

[4]The constant lower bound on σ_{0n}^2 can be relaxed as long as $s_{\max}(\Sigma_{0n})/s_{\min}(\Sigma_{0n}) \lesssim n$.

for all $j \geq j_0$ for some $j_0 \in \mathbb{N}$. Additionally, assume

$$\pi_k(k = k_{0n}) \geq \exp(-C s_n k_{0n} \log n), \tag{4.2}$$

where s_n is the sequence appearing in (A2). For instance, a Poisson distribution on k with rate parameter 1 will satisfy (4.1) and (4.2) if $\log k_{0n} \leq s_n \log n$, which is automatically satisfied given (A1).

Conditional on k, we consider two classes of prior distributions on the factor loadings Λ_n. We first consider a class of point mass mixture priors on the loadings similar to that advocated by [41],

$$\begin{aligned} &\lambda_{jh} \mid k, \qquad \pi \sim (1-\pi)\delta_0 + \pi g(\cdot), \qquad j = 1, \ldots, p_n; h = 1, \ldots, k, \\ &\qquad\qquad k \sim \pi_k, \qquad \pi \sim \mathrm{Beta}(1, \kappa k_{0n} p_n + 1), \qquad \kappa > 0, \end{aligned} \tag{PL1}$$

where δ_0 denotes a point mass at zero and g is an absolutely continuous density on $\mathbb{R}$ with exponential tails or heavier.

For linear models, [38] showed that such point mass mixture priors with a beta hyper-prior on the mixture probability lead to an automatic multiplicity correction. [29] proved optimality results in estimating the predictive distribution under such priors in generalized linear models accommodating diverging numbers of predictors. Castillo and van der Vaart [17] studied concentration properties of a class of prior distributions similar to (PL1) on a high-dimensional normal mean and showed that they lead to the minimax optimal rate of convergence.

As mentioned in the Introduction, although point mass mixture priors are conceptually appealing in allowing exact sparsity and often leading to appealing theoretical properties, posterior computation under such priors can be daunting in high-dimensional cases. As an alternative, a rich variety of continuous shrinkage priors have been developed that admit a scale mixture representation [36]. A fundamental hurdle in studying theoretical properties of such priors is the difficulty of obtaining tight bounds on their concentration and implied dimensionality. With the motivation of developing a continuous shrinkage prior that can be shown to concentrate near sparse vectors and approximate point mass mixture priors, we propose a novel class of priors. We use such priors for the factor loadings, but they should be broadly applicable in other high-dimensional settings.

Let $\mathrm{DE}(\psi)$ denote the Laplace or double-exponential density with scale parameter ψ with a density given by

$$f(x) = \frac{1}{2\psi} e^{-|x|/\psi}, \qquad x \in \mathbb{R}. \tag{4.3}$$

Draw the elements of a high-dimensional vector $\theta \in \mathbb{R}^p$ through the following hierarchical mechanism:

$$\theta_j \sim \mathrm{DE}(\tau\gamma_j), \qquad \tau \sim f_\tau, \gamma = (\gamma_1, \ldots, \gamma_p)' \sim f_\gamma. \tag{PS}$$

In (PS), $\tau > 0$ is a global scale parameter and $\gamma \in \Delta^{p-1}$ is a vector of local scale parameters. We set f_τ to be an exp(1/2) density. We draw $\tilde{\gamma} = (\gamma_1, \ldots, \gamma_{p-1})^{\mathrm{T}} \in \Delta_0^{p-1}$ from a $\mathrm{Dir}(\alpha/p, \ldots, \alpha/p)$ density and set $\gamma_p = 1 - \sum_{j=1}^{p-1} \gamma_j$. For a detailed discussion on the properties of the prior (PS), refer to Section 7.

Given k, we consider the prior (PS) on the vectorized loadings $\mathrm{vec}(\Lambda_n) \in \mathbb{R}^{p_n k}$ as an alternative to (PL1); note that the Dirichlet concentration parameter becomes $\alpha/(p_n k)$ in this case.

5. Main results. With the prior specification complete, we now state the main results of this paper. The proofs are available in Section 9. Theorems 5.1 and 5.2 assume the true number of factors to be bounded, which is generalized in Theorem 5.3. Recall the class of "true" covariance matrices $\mathcal{C}_{0n}$ from Section 3. We first establish the rate of posterior convergence in operator norm using the point mass priors (PL1) on the loadings in Theorem 5.1.

THEOREM 5.1. *Suppose $\Sigma_{0n} \in \mathcal{C}_{0n}$ with $s_n \gtrsim \log p_n$ and $k_{0n} = O(1)$, and model (2.1) is fitted with a prior distribution on the number of factors satisfying (4.1) and (4.2). Assume independent priors $\Pi(\Lambda \mid k)$ and $\Pi(\sigma^2)$ on the loadings and the residual variances as in (PL1) and (PR), respectively. Then, with $\varepsilon_n = c_n \sqrt{\frac{s_n \log p_n}{n}} \sqrt{\log n}$ and for some constant $M > 0$,*

$$\lim_{n \to \infty} \mathbb{E}_{\Sigma_{0n}} \Pi_n(\|\Sigma_n - \Sigma_{0n}\|_2 > M\varepsilon_n \mid \mathbf{y}^{(n)}) = 0, \tag{5.1}$$

where $\mathbb{E}_{\Sigma_{0n}}$ denotes an expectation with respect to the joint distribution of $\mathbf{y}^{(n)}$.

We next show in Theorem 5.2 that our proposed shrinkage prior on the loadings achieves the same posterior rate of convergence as for the point mass mixture priors.

THEOREM 5.2. *Assume the same setup as in Theorem 5.1, with the point mass prior (PL1) on the loadings replaced by the shrinkage prior (PS) on the vectorized loadings given k. Then (5.1) is satisfied with $\varepsilon_n = c_n \sqrt{\frac{s_n \log p_n}{n}} \sqrt{\log n}$.*

We show in Section 5.1 that $c_n \sqrt{s_n \log p_n / n}$ is the minimax rate of estimating $\Sigma_{0n} \in \mathcal{C}_{0n}$ in operator norm with $k_{0n} = O(1)$. Thus, the posterior rate of convergence obtained in both Theorems 5.1 and 5.2 is equal to the minimax rate up to a $\sqrt{\log n}$ term. For a general k_{0n}, we establish analogous versions of Theorems 5.1 and 5.2 below.

THEOREM 5.3. *If $\Sigma_{0n} \in \mathcal{C}_{0n}$ with $s_n k_{0n} \gtrsim \log p_n$, the convergence rates in both Theorems 5.1 and 5.2 are modified to $c_n k_{0n}^{3/2} \sqrt{\frac{s_n \log p_n}{n}} \sqrt{\log n}$.*

Clearly, Theorem 5.3 permits consistent estimation in operator norm even if $p_n = \exp(n^\alpha)$, $s_n = n^\beta$ and $k_{0n} = n^\gamma$ for appropriate $\alpha, \beta, \gamma > 0$. At this point, we do not know whether the rate obtained in Theorem 5.3 is minimax-optimal and substantial further work seems necessary to prove such a result.

5.1. *A lower bound to the minimax rate.* Minimax optimal rates in operator norm for high-dimensional covariance matrix estimation have been established for a class of bandable matrices by [13] and a class of covariance matrices with sparse columns by [14]. Although $\mathcal{C}_{0n}$ has a nonempty intersection with the class $\mathcal{G}_0(c_{n,p})$ in [14], there exists a large subclass of matrices which lie in one and not in the other. Moreover, the existing minimax results on large covariance estimation assume the eigenvalues of the true sequence of covariance matrices to be bounded. For example, [13] and [14] assume that y_i is sub-Gaussian, that is, for all $t > 0$ and $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$, $\mathbb{P}(|v^{\mathrm{T}}(y_1 - \mathbb{E}y_1)| > t) \leq \exp(-t^2/2\tau^2)$. The parameter τ is assumed to be a constant and its role in the rate is not characterized. For $y_1 \sim \mathrm{N}_p(0, \Sigma)$, a standard tail bound for the normal distribution implies

$$\mathbb{P}(|v^{\mathrm{T}} y_1| > t) \leq \exp\left(-\frac{t^2}{2v^{\mathrm{T}}\Sigma v}\right) \leq \exp\left(-\frac{t^2}{2\|\Sigma\|_2}\right).$$

For $\Sigma = \Lambda\Lambda^{\mathrm{T}} + \sigma^2 \mathrm{I}_p \in \mathcal{C}_{0n}, \|\Sigma\|_2 = \|\Lambda\|_2^2 + \sigma^2 \asymp c_n$ by Assumption 3.2, so that $\tau \asymp \sqrt{c_n}$ in our case. Hence, the growth rate of $\|\Sigma\|_2$ needs to be accounted for in our calculations. With this motivation, we study minimax lower bounds for $\mathcal{C}_{0n}$ in Theorem 5.4 below.

THEOREM 5.4. *If $\hat{\Sigma}_n$ is a sequence of estimators of $\Sigma_{0n} \in \mathcal{C}_{0n}$ with $k_{0n} = O(1)$, then*

$$\inf_{\hat{\Sigma}_n} \sup_{\Sigma_{0n} \in \mathcal{C}_{0n}} \|\hat{\Sigma}_n - \Sigma_{0n}\|_2 \geq c_n \sqrt{s_n \frac{\log p_n}{n}}. \tag{5.2}$$

PROOF. We will use Fano's lemma to derive a lower bound for the minimax risk. Let $\mathcal{F}$ be a parameter space of covariance matrices and we observe $y_1, \ldots, y_n \sim \mathrm{N}(0, \Sigma)$ with $\Sigma \in \mathcal{F}$. Let $\Theta = \{\Sigma_{(1)}, \ldots, \Sigma_{(m_n)}\}, m_n \geq 2$ be a finite subset of $\mathcal{F}$ and let $\mathbb{P}^{(j)}$ denote the joint distribution of $y_1, \ldots, y_n$ independently distributed as $\mathrm{N}(0, \Sigma_{(j)})$, $1 \leq j \leq m_n$. Let $\hat{\Sigma}$ be an estimator for Σ. Suppose for all $j \neq j'$, we have that

$$d(\Sigma_{(j)}, \Sigma_{(j')}) \geq d_{m_n}, \qquad \mathrm{KL}(\mathbb{P}^{(j)}, \mathbb{P}^{(j')}) \leq K_{m_n}.$$

Letting $\mathbb{E}_j$ denote the expectation under $\mathbb{P}^{(j)}$, Fano's lemma (as in [42]) implies

$$\max_{1 \leq j \leq m_n} \mathbb{E}_j d(\hat{\Sigma}, \Sigma) \geq \frac{d_{m_n}}{2}\left(1 - \frac{K_{m_n} + \log 2}{\log m_n}\right). \tag{5.3}$$

We first introduce notation and then proceed to construct our finite parameter set Θ. Let $q_n = p_n - 1$. Define $\mathcal{M} := \{x \in \mathbb{R}^{q_n} : x_j \in \{0,1\}\ \forall j, \|x\|_1 = s_n\}$ to be the collection of all binary vectors of length q_n with exactly s_n ones. Let $\|\cdot\|_H$ denote the Hamming distance between two binary strings, so that $\|x-y\|_H = \sum_{j=1}^{q_n} 1(x_j \neq y_j)$. Let $b_j = (\theta_j, 0)$ denote the p_n-dimensional vector obtained by appending zero at the end of θ_j. With this notation, set

$$\Sigma_{(j)} = \beta \mathrm{I}_{p_n} + \gamma b_j b_j^{\mathrm{T}} + \kappa e_{p_n} e_{p_n}^{\mathrm{T}}, \qquad j \in \mathcal{M},$$

where $e_p \in \mathbb{R}^p$ is the vector with 1 in the pth coordinate and zero elsewhere, and $\gamma < \beta < \kappa$ are sequences to be chosen below.

We now state Lemmas 5.5 and 5.6; for clarity in notation, we drop the subscript n in both lemmata. Refer to the Appendix for a proof.

LEMMA 5.5. *For $j \neq j'$ and $0 \leq r \leq s$, if $\|\theta_j - \theta_{j'}\|_H = 2(s-r)$, then*

$$\|\Sigma_{(j)} - \Sigma_{(j')}\|_2 = \gamma\sqrt{s^2 - r^2}, \qquad \mathrm{KL}(\mathbb{P}^{(j)}, \mathbb{P}^{(j')}) = \frac{n}{2}\frac{t^2}{ts+1}(s^2 - r^2),$$

where $t = \gamma/\beta$.

LEMMA 5.6. *Given $s \geq 6$, there exists a subset $\mathcal{M}_0 = \{\theta_1, \ldots, \theta_m\}$ of $\mathcal{M}$ with $m \asymp \exp(Cs\log p)$ and $\|\theta_j - \theta_{j'}\|_H \geq s/3$ for all $1 \leq j \neq j' \leq m$, where C is a positive constant independent of p.*

We set $\Theta = \{\Sigma_{(j)} : \theta_j \in \mathcal{M}_0\}$. Since $\|\theta_j - \theta_{j'}\|_H \geq s_n/3$ by Lemma 5.6, the quantity $r_n = r_n(j, j')$ appearing in Lemma 5.5 is bounded above by $5s_n/6$ for all pairs $j \neq j' \in \mathcal{M}_0$. Hence, we can choose $d_{m_n} = C_1\gamma s_n$ and $K_{m_n} = n(ts_n)^2 = n(\gamma s_n/\beta)^2$ in (5.3). To obtain d_{m_n} as a lower bound to the minimax risk up to a constant, we need to set $K_{m_n}/\log m_n = C'$ for some constant $C' \in (0,1)$. Since $\log m_n \asymp Cs_n \log p_n$, we obtain, by choosing $\beta = c_n$, that $d_{m_n}^2 = C(\gamma s_n)^2 = C\frac{c_n^2 s_n \log p_n}{n}$ for some absolute constant C. $\square$

6. Simulation studies. In this section, we consider a number of simulation cases to compare our proposed continuous shrinkage prior (PS) with existing methods including the point mass priors (PL1) on the loadings matrix and the sample covariance matrix $S = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}$. For prior (PL1), we use a standard Laplace distribution on the signal coefficients.

We also compare our methods with Principal Orthogonal complement Thresholding (POET) of [21] which is based on an additive decomposition of the covariance matrix in terms of a low rank matrix and a sparse residual covariance matrix. POET estimates the factors and the loadings by thresholding the principal components of the sample covariance matrix. Finally,

Table 1

Simulation setting 1. Comparative performance in covariance matrix estimation for (PL1), (PS), POET, AT. The average error in operator norm across simulation replicates is tabulated

n	**50**				**100**			
p_n	**100**		**200**		**100**		**200**	
k_{0n}	**1**	$\log p_n$	**1**	$\log p_n$	**1**	$\log p_n$	**1**	$\log p_n$
(PL1)	0.98 (0.43)	2.84 (1.12)	10.06 (5.68)	9.79 (4.90)	8.83 (0.12)	12.82 (1.42)	15.90 (0.26)	16.07 (1.77)
(PS)	1.03 (0.38)	3.95 (1.69)	5.96 (1.81)	7.01 (2.01)	1.74 (0.83)	3.43 (1.10)	3.66 (1.83)	4.21 (1.20)
POET	2.89 (0.41)	6.98 (1.28)	8.90 (2.11)	12.41 (2.69)	3.08 (0.64)	5.72 (1.09)	7.32 (1.51)	7.51 (1.44)
AT	1.93 (0.57)	4.71 (2.97)	6.92 (5.43)	8.86 (3.79)	2.11 (0.71)	3.26 (1.08)	3.80 (2.03)	4.37 (1.34)
SC	2.79 (0.36)	7.08 (1.33)	9.01 (2.22)	12.73 (2.80)	3.06 (0.65)	5.73 (1.14)	7.34 (1.52)	7.52 (1.46)

we provide results for the adaptive thresholding method (AT) of [12] which thresholds the entries of the sample covariance matrix, with the resulting thresholded estimator $\hat{\Sigma}$ being of the form $\hat{\Sigma}_{jj'} = S_{jj'}1(|S_{jj'}| > \delta\kappa_{jj'})$, where δ is a tuning parameter and $\kappa_{jj'}$ is a threshold specific to the corresponding entry of S. We chose the tuning parameter δ by 5-fold cross-validation as suggested by [12]. We also implemented the same procedure with the default choice of $\delta = 2$; the results were worse in all cases, and hence are not reported.

We describe the two simulation settings below:

1. $y_i, i = 1, \ldots, n$ are generated from $\mathrm{N}_{p_n}(0, \Sigma_{0n})$, where $\Sigma_{0n} = \Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} + \sigma_0^2 \mathrm{I}_{p_n}$ and Λ_{0n} is a $p_n \times k_{0n}$ matrix with $s_n = \log p_n$ nonzero entries per column and $k_{0n} = 1$ or $\log p_n$. The nonzero entries were drawn uniformly between 1 and 2. These simulations were designed to mimic assumptions (A0)–(A4) in Section 3.
2. This setting is designed to illustrate the performance of our method under model misspecification. We let $\Sigma_{0n} = \Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} + \Omega_{0n}$, where Λ_{0n} is as in simulation setting (1), but Ω_{0n} is nondiagonal, corresponding to the covariance matrix of an autoregressive sequence with pure error variance 0.4 and autoregressive coefficient 0.1.

For each simulation setting, we choose two sample sizes, namely $n = 50, 100$ and for each value of n, we let $p_n = 100, 200$. For each (n, p_n) pair, we consider 50 simulation replicates. For the Bayesian methods, the posterior mean is used as a point estimate. Tables 1 and 2 summarize the results across the simulation replicates for the two simulation settings, respectively, to compare the operator norm difference between the estimator resulting from the different methods and the truth. In particular, the average error across 50 replicates is provided, with standard error in parenthesis.

The results for (PS) and (PL1) were reported based on 10,000 runs of the Gibbs sampler with 5000 burn-in. From Tables 1 and 2, it becomes evident

TABLE 2

Simulation Setting 2. Comparative performance in covariance matrix estimation for (PL1), (PS), POET, AT. The average error in operator norm across simulation replicates is tabulated

n	**50**				**100**			
p_n	**100**		**200**		**100**		**200**	
k_{0n}	**1**	$\log p_n$	**1**	$\log p_n$	**1**	$\log p_n$	**1**	$\log p_n$
(PL1)	1.73 (1.26)	5.30 (3.92)	11.92 (2.82)	13.41 (4.03)	17.01 (0.22)	19.37 (1.74)	10.04 (0.08)	22.10 (0.52)
(PS)	2.44 (1.40)	5.42 (2.67)	4.12 (2.86)	7.98 (3.23)	2.01 (1.44)	4.56 (1.49)	2.04 (1.12)	5.23 (2.10)
POET	3.59 (0.84)	7.16 (1.84)	7.14 (1.59)	12.63 (2.89)	3.93 (1.17)	7.39 (1.69)	3.90 (0.71)	10.13 (2.09)
AT	2.32 (1.49)	5.50 (3.09)	4.04 (2.99)	8.26 (4.16)	2.12 (1.62)	4.45 (1.63)	1.97 (0.90)	4.96 (2.28)
SC	3.63 (0.88)	7.32 (1.95)	7.26 (1.66)	12.85 (3.07)	3.95 (1.19)	7.44 (1.75)	3.88 (0.72)	10.24 (0.26)

that when the number of model parameters increase, the performance of (PL1) deteriorates due to possibly slower convergence of the MCMC, while (PS) has more robust performance. Even in Table 2, where the truth is misspecified for both (PS) and AT, and in fact designed to favor POET, (PS) performs at least equally or better than the competitors. For each MCMC iteration, the runtime for (PS) scaled approximately linearly with n and p, though we are not aware of sharp theoretical bounds on MCMC convergence in high dimensions guaranteeing polynomial time convergence unlike many frequentist estimators.

7. Some properties of shrinkage priors in high-dimensional settings. We develop a number of properties of the proposed shrinkage prior (PS) in high-dimensional settings; the results are used to prove the main results on posterior concentration, but are also of independent interest. Proofs of all the results are deferred to the Appendix.

Let θ be a p-dimensional vector and $\theta_0 \in l_0[s;p]$ be an s-sparse vector with $s \ll p$. Depending on the problem, θ might correspond to a high-dimensional mean vector, a vector of regression coefficients or a column of the factor loadings, with θ_0 corresponding to a sparse truth.[5] A quantity of fundamental importance in studying the behavior of the posterior distribution in high-dimensional problems is the *prior concentration* around an arbitrary sparse vector θ_0, which is defined as the noncentered small ball probability

$$\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon), \tag{7.1}$$

for ε small. It can be shown that if θ_j's are i.i.d. standard normal,

$$\sup_{\theta_0 \in l_0[s;p]} \mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon) \leq e^{-Cp \log(1/\varepsilon)},$$

[5]For us, θ and θ_0 correspond to the vectorized loadings Λ_n and Λ_{0n}, respectively.

which decays exponentially with p for fixed s limiting the ability of the posterior to concentrate on sparse θ_0. However, with appropriate point mass mixture priors having a probability mass at zero and $\|\theta_0\|$ bounded, the small ball probability (7.1) can be improved to $e^{-Cs\log(1/\varepsilon)}$ [17].

For reasons mentioned in Section 4, there has been a recent thrust on developing *one-group* alternatives to the *two-group* mixture priors using continuous shrinkage priors, which can be often represented as a *global–local* scale mixture [36] of Gaussians. Despite computational advantages with this family of shrinkage priors, their concentration properties are understudied. Our proposed prior (PS), which can be expressed as a Gaussian scale mixture, favors a large subset of the θ_j to be *simultaneously* close to zero while inflicting minimal shrinkage on the rest, and thus achieve a concentration similar to point mass mixture priors. In the following Lemma 7.1, we present a *nonasymptotic* bound to the prior concentration for (PS).

LEMMA 7.1. *Suppose $\theta \sim$ (PS). Let $\theta_0 \in l_0[s;p], 1 \le s \le p$ and $s/p \le 1/2$. Then, for any $\varepsilon \in (0,1)$ small enough,*

$$\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon) \ge \exp[-C \max\{\|\theta_0\|_2^2, s\log(s/\varepsilon), \log p\}]$$

for some constant $C > 0$.

We also state an auxiliary Lemma 7.2 which is used to prove Lemma 7.1; refer to the supplemental document for a proof.

LEMMA 7.2. *Let $\eta \in \mathbb{R}^s$ denote a random vector with independent components $\eta_j \sim \mathrm{DE}(\psi_j)$. If there exist numbers $a, b > 0$, such that $\psi_j \in [a, b]$ for all $j = 1, \dots, s$, then for any $\delta > 0$ and $\eta_0 \in \mathbb{R}^s$,*

$$\mathbb{P}(\|\eta - \eta_0\|_2 < \delta) \ge \exp\left\{-\frac{C_1}{a^2}\sum_{j=1}^{s}|\eta_{0j}|^2 - C_2 s - s|\log\{\delta/(b\sqrt{s})\}|\right\}$$

for constant $C_1, C_2 > 0$.

We next show that the shrinkage prior (PS) does not spread its mass across too many dimensions. A point mass mixture prior allows a high-dimensional vector to collapse onto fewer dimensions and the implied dimensionality can be naturally studied through appropriate tail bounds for the induced prior on $|\operatorname{supp}(\theta)|$, which is a random variable supported on $\{0, 1, \dots, p\}$. Such bounds on the prior dimensionality are useful to control the posterior model size [17]. However, continuous shrinkage priors do not allow exact zeroes in θ and clearly $\mathbb{P}(|\operatorname{supp}(\theta)| = p) = 1$. We instead use a generalized definition of the support of a vector as the subset of entries which are larger than a small number δ in magnitude. For any $\delta > 0$, we denote the corresponding subset to be $\operatorname{supp}_\delta(\theta)$, so that $\operatorname{supp}_\delta(\theta) = \{j : |\theta_j| > \delta\}$.

In the following Lemma 7.3, we provide a *nonasymptotic* tail bound for $|\operatorname{supp}_\delta(\theta)|$, the *number* of entries in θ larger than δ in magnitude.

LEMMA 7.3. *Let $\varepsilon \in (0,1)$ and $\delta = \varepsilon/p$ with $\varepsilon > 1/p^B$ for some $B > 0$. If θ is drawn according to the prior (PS) and $s \gtrsim \log p$, then there exists a constant $A > 0$ such that*

$$\mathbb{P}(|\operatorname{supp}_\delta(\theta)| > As) \le e^{-Cs}$$

for some constant $C > 0$. Moreover, the constant C appearing in the exponent can be made arbitrarily large by choosing A large enough.

A final important property of (PS) is established through the following deviation result on the l_1 norm of θ.

LEMMA 7.4. *For $t \ge 1$, $\mathbb{P}[\|\theta\|_1 \ge t] \le 2e^{-C\sqrt{t}}$.*

8. Construction of test functions. An important step [24] in Bayesian asymptotic theory for establishing posterior contraction rates is to develop a test function for the true parameter versus the complement of a ball of radius $\varepsilon > 0$ (in an appropriate norm) around the truth with type-I and II error rates of the order $\exp(-Cn\varepsilon^2)$. Under the Hellinger or L_1 distance between densities, existence of such tests is guaranteed by the seminal work of [9] and [32]; the same is true for norms compatible to the above norms [25]. However, when the object of interest is *not the density itself*, but rather some high-dimensional parameter indexing the density with a norm of discrepancy relevant to the space the parameter lives in, the test arising from Birgé–Le Cam theory might fail to produce the desired error rates in the *norm of interest*.

In the context of nonparametric function estimation in general L_r norms, [26] advocated using concentration inequalities based on empirical process techniques as an alternative to the traditional testing framework. Castillo and Van Der Vaart [17] used deviation bounds for the likelihood ratio test in estimating a high dimensional mean in *Euclidean norm*. An important contribution of the present paper is to utilize recently developed concentration results for random (self-adjoint) matrices [39, 40] to devise a test function.

Using a version of the matrix Bernstein inequality (Theorem 6.2 in [39]), it can be shown that the sample estimator $\hat{\Sigma} = n^{-1}\sum_{i=1}^n y_i y_i^{\mathrm{T}}$ has appropriate concentration around $\mathbb{E}\hat{\Sigma} = \Sigma_0$ when the "effective rank" $r_e(\Sigma_0) := \operatorname{tr}(\Sigma_0)/\|\Sigma_0\|_2$ is modest compared to p [11, 40]. However, for $\Sigma_0 = \Lambda_0\Lambda_0^{\mathrm{T}} + \sigma_0^2 \mathrm{I}_p \in \mathcal{C}_{0n}$, $r_e(\Sigma_0)$ can scale in the order of p, prohibiting us from using $\hat{\Sigma}$ as an estimator to construct the test. A crucial observation is that even if $\Sigma_0 \in \mathcal{C}_{0n}$ does not necessarily have a small effective rank, the larger contribution to the operator norm of Σ_0 ($\|\Sigma_0\|_2 = \|\Lambda_0\|_2^2 + \sigma_0^2$) comes from the low rank part by (A3). We exploit this to design a novel projection based test in Theorem 8.1 below, where the types I and II error rates can be expressed in terms of deviation bounds of a $k_{0n} \times k_{0n}$ sample covariance matrix from its mean. Dependence of all quantities on n has been made explicit from this point onwards.

THEOREM 8.1. *Recall the sequences k_{0n} and c_n from Assumptions 3.1 and 3.2, respectively. Let $\Sigma_{0n} \in \mathcal{C}_{0n}$ with the corresponding $\Lambda_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$. Let $B_{j,n} = \{\Sigma_n \in \mathcal{C}_n : j\varepsilon_n \le \|\Sigma_n - \Sigma_{0n}\|_2 < (j+1)\varepsilon_n\}$ denote an annulus of inner radius $j\varepsilon_n$ and outer radius $(j+1)\varepsilon_n$ in operator norm around Σ_{0n} for some integer $j > 1$ and sequence $\varepsilon_n > 0$. Assume $\varepsilon_n \ge c_n k_{0n}\sqrt{\frac{\log k_{0n}}{n}}$ and $\varepsilon_n \log j \le c_n$. Fix $\Sigma_{1n} \in B_{j,n}$ and let $E_{j,n} = \{\Sigma_n \in B_{j,n} : \|\Sigma_n - \Sigma_{1n}\|_2 < j\varepsilon_n/2\}$ denote an operator norm ball in $B_{j,n}$ around Σ_{1n} of radius $j\varepsilon_n/2$.*

Based on n i.i.d. samples $y_1, \dots, y_n$ from $\mathrm{N}_{p_n}(0, \Sigma_n)$, consider testing the point null vs. composite alternative hypothesis

$$H_0 : \Sigma_n = \Sigma_{0n} \quad \text{versus} \quad H_1 : \Sigma_n \in E_{j,n}. \tag{8.1}$$

Define $x_i = (1/c_n)\Lambda_{0n}^{\mathrm{T}} y_i$ and $z_i = \Lambda_{0n} x_i$ for $i = 1, \dots, n$, so that $x_i \in \mathbb{R}^{k_{0n}}$ and $z_i \in \mathbb{R}^{p_n}$. Let $\hat\Sigma_y = n^{-1}\sum_{i=1}^n y_i y_i^{\mathrm{T}}$ and define $\hat\Sigma_x = (1/c_n^2)\Lambda_{0n}^{\mathrm{T}}\hat\Sigma_y\Lambda_{0n}, \hat\Sigma_z = \Lambda_{0n}\hat\Sigma_x\Lambda_{0n}^{\mathrm{T}}$. Let $\phi_{j,n}$ denote a test function for (8.1) defined as

$$\phi_{j,n} = 1_{\{\|\hat\Sigma_z - \Sigma_{0n}\|_2 \ge j\varepsilon_n/4\}}. \tag{8.2}$$

Then, the type-I and type-II error rates of $\phi_{j,n}$ satisfy:

$$\mathbb{E}_0\phi_{j,n} \le \exp\left\{-\frac{Cnj^2\varepsilon_n^2}{c_n^2 k_{0n}^2}\right\}, \tag{8.3}$$

$$\sup_{\Sigma_n \in E_{j,n}} \mathbb{E}_{\Sigma_n}(1 - \phi_{j,n}) \le \exp\left\{-\frac{Cn(\log j)^2\varepsilon_n^2}{c_n^2 k_{0n}^2}\right\} \tag{8.4}$$

for some constant $C > 0$, where $\mathbb{E}_{\Sigma_n}$ denotes an expectation under the distribution of $\mathbf{y}^{(n)}$ under $\mathrm{N}(0, \Sigma_n)$ and $\mathbb{E}_0$ is a shorthand for $\mathbb{E}_{\Sigma_{0n}}$.

REMARK. If the condition $\varepsilon_n \log j \le c_n$ is replaced by $j^\delta\varepsilon_n \le c_n$ for some $0 < \delta \le 1$, the type-II error bound in (8.4) becomes $\exp\{-Cnj^{2\delta}\varepsilon_n^2/(c_n^2 k_{0n}^2)\}$.

PROOF OF THEOREM 8.1. We shall make use of a matrix concentration result from [11]. Let $u_1, \dots, u_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_q(0, \Sigma)$ and $\hat\Sigma := n^{-1}\sum_{i=1}^n u_i u_i^{\mathrm{T}}$ denote the sample covariance matrix. Proposition A.4 in [11] implies that for any $s > 0$ such that $s + \log q < n$,

$$\mathbb{P}\left[\|\hat\Sigma - \Sigma\|_2 > C\,\mathrm{tr}(\Sigma)\sqrt{\frac{s + \log q}{n}}\right] \le e^{-s}. \tag{8.5}$$

We adapt a fact from Lemma 5.36 of [40]. For a $p \times k$ matrix B with $p > k$, suppose $\|B^{\mathrm{T}}B - \mathrm{I}_k\|_2 \le \max\{\delta, \delta^2\}$ for some $\delta > 0$. Then

$$1 - \delta \le s_{\min}(B) \le s_{\max}(B) \le 1 + \delta. \tag{8.6}$$

Finally, we index matrices that appear frequently in the sequel. Define

$$\begin{aligned} G_n &:= \frac{1}{c_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n} - \mathrm{I}_{k_{0n}}, \qquad \Psi_n := \frac{1}{c_n}\Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} - \mathrm{I}_{p_n}, \\ \Gamma_n &:= \left(1 + \frac{\sigma_{0n}^2}{c_n}\right)\mathrm{I}_{k_{0n}}. \end{aligned} \tag{8.7}$$

Note that G_n is the matrix appearing in (A3). The nonzero eigenvalues of G_n and Ψ_n are the same; hence, $\|G_n\|_2 = \|\Psi_n\|_2$.

Type-I error: Recall $\hat{\Sigma}_z$ and $\hat{\Sigma}_x$ from the theorem statement. We proceed to bound $\mathbb{E}_0\phi_{j,n} = \mathbb{P}_0[\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 \ge j\varepsilon_n/4]$. By the triangle inequality and Lemma 1.1 in the supplemental document,

$$\begin{aligned} &\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 \\ &\quad \le \left\|\Lambda_{0n}\hat{\Sigma}_x\Lambda_{0n}^{\mathrm{T}} - \Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} - \frac{\sigma_{0n}^2}{c_n}\Lambda_{0n}\Lambda_{0n}^{\mathrm{T}}\right\|_2 + \sigma_{0n}^2\|\Psi_n\|_2 \\ &\quad \le \|\Lambda_{0n}\|_2^2\|\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2 + \|\Lambda_{0n}\|_2^2\|\mathbb{E}_0\hat{\Sigma}_x - \Gamma_n\|_2 + \sigma_{0n}^2\|G_n\|_2. \end{aligned} \tag{8.8}$$

A simple calculation yields

$$\begin{aligned} \mathbb{E}_0\hat{\Sigma}_x &= \frac{1}{c_n^2}\Lambda_{0n}^{\mathrm{T}}[\Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} + \sigma_{0n}^2\mathrm{I}_{p_n}]\Lambda_{0n} \\ &= \left(\frac{1}{c_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}\right)^2 + \frac{\sigma_{0n}^2}{c_n}\frac{1}{c_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}. \end{aligned} \tag{8.9}$$

Substituting this in (8.8) and using triangle inequality, the sum of the second and third term in (8.8) can be bounded above by

$$\|\Lambda_{0n}\|_2^2\left\|\left(\frac{1}{c_n}\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}\right)^2 - \mathrm{I}_{k_{0n}}\right\|_2 + \sigma_{0n}^2(1/c_n + 1)\|G_n\|_2.$$

Recall $\|G_n\|_2 = o(k_{0n}\sqrt{\log k_{0n}/n})$ by (A3). In addition, $\|\Lambda_{0n}\|_2 \le 2\sqrt{c_n}$ by (A3) and $\sigma_{0n}^2 \le c_n$ from (A4). Note that $\|A - \mathrm{I}_{k_{0n}}\|_2 < \delta$ for A symmetric and some $\delta \in (0,1)$ implies that $\|A^2 - \mathrm{I}_{k_{0n}}\|_2 \le 3\delta$. Using these facts, the expression in the above display can be bounded above by $(13c_n + 1)\|G_n\|_2$. Since we have assumed $\varepsilon_n \ge c_n k_{0n}\sqrt{\frac{\log k_{0n}}{n}}$ in the condition of the theorem, $(13c_n + 1)\|G_n\|_2$ can be bounded above by $j\varepsilon_n/8$ for n large enough. Substituting this bound in (8.8),

$$\mathbb{P}_0[\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 \ge j\varepsilon_n/4] \le \mathbb{P}_0[\|\Lambda_{0n}\|_2^2\|\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2 \ge j\varepsilon_n/8].$$

Using $\|\Lambda_{0n}\|_2^2 \lesssim c_n$ one more time, we have

$$\mathbb{E}_0\phi_{j,n} \leq \mathbb{P}_0[\|\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2 \geq Cj\varepsilon_n/c_n]. \tag{8.10}$$

By definition, $\hat{\Sigma}_x = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{T}}$ is a $k_{0n} \times k_{0n}$ sample covariance matrix. We now invoke (8.5) to bound the deviation of $\hat{\Sigma}_x$ from its expectation under $\mathbb{P}_0$ in (8.10). From (8.9) and using $\mathrm{tr}(A^2) = \|A\|_F^2$ for A symmetric, $\mathrm{tr}(\mathbb{E}_0\hat{\Sigma}_x) = \|\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}/c_n\|_F^2 + (\sigma_{0n}^2/c_n)\|\Lambda_{0n}/\sqrt{c_n}\|_F^2$. Recall $\sigma_{0n}^2 \leq c_n$ by (A4). By Lemma 1.1 in the supplemental document, $\|\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}/c_n\|_F \leq \|\Lambda_{0n}/\sqrt{c_n}\|_F\|\Lambda_{0n}/\sqrt{c_n}\|_2 \leq 2\|\Lambda_{0n}/\sqrt{c_n}\|_F$. Hence, $\mathrm{tr}(\mathbb{E}_0\hat{\Sigma}_x) \leq 5\|\Lambda_{0n}/\sqrt{c_n}\|_F^2$. Using $\|\Lambda_{0n}\|_F \leq \sqrt{k_{0n}}\|\Lambda_{0n}\|_2 \leq 2\sqrt{c_n k_{0n}}$, $\mathrm{tr}(\mathbb{E}_0\hat{\Sigma}_x)$ can be bounded above by Ck_{0n}.

Choose $s = Cnj^2\varepsilon_n^2/(k_{0n}^2c_n^2)$. Since $\varepsilon_n \geq c_n k_{0n}\sqrt{\log k_{0n}/n}$, we have $s \gtrsim \log k_{0n}$ and hence $C\,\mathrm{tr}(\mathbb{E}_0\hat{\Sigma}_x)\sqrt{(s+\log k_{0n})/n} \lesssim C\,\mathrm{tr}(\mathbb{E}_0\hat{\Sigma}_x)\sqrt{s/n} \leq Cj\varepsilon_n/c_n$. By (8.5), the expression in the right-hand side of (8.10) is then bounded above by $e^{-s} = \exp\{-Cnj^2\varepsilon_n^2/(k_{0n}^2c_n^2)\}$, proving (8.3).

Type-II error: Fix $\Sigma_n \in E_{j,n}$. We proceed to bound $\mathbb{E}_{\Sigma_n}(1-\phi_{j,n}) = \mathbb{P}_{\Sigma_n}[\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 < j\varepsilon_n/4]$. By repeatedly using the triangle inequality, we obtain

$$\begin{aligned}\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 &\geq \|\Lambda_{0n}\|_2^2\left\|\hat{\Sigma}_x - \mathrm{I}_{k_{0n}} - \frac{\sigma_{0n}^2}{c_n}\mathrm{I}_{k_{0n}}\right\|_2 - \sigma_{0n}^2\left\|\frac{1}{c_n}\Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} - \mathrm{I}_{p_n}\right\|_2 \\ &\geq \|\Lambda_{0n}\|_2^2\{\|\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \Gamma_n\|_2 - \|\hat{\Sigma}_x - \mathbb{E}_{\Sigma_n}\hat{\Sigma}_x\|_2\} - \sigma_{0n}^2\|\Psi_n\|_2.\end{aligned}$$

Recall $\|\Psi_n\|_2 = \|G_n\|_2$. Therefore, on the set $\{\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 < j\varepsilon_n/4\}$,

$$\begin{aligned}\|\Lambda_{0n}\|_2^2\|\hat{\Sigma}_x - \mathbb{E}_{\Sigma_n}\hat{\Sigma}_x\|_2 &\geq \|\Lambda_{0n}\|_2^2\|\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \Gamma_n\|_2 - \sigma_{0n}^2\|G_n\|_2 - \frac{j\varepsilon_n}{4} \\ &\geq \|\Lambda_{0n}\|_2^2\|\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2 \\ &\quad - \|\Lambda_{0n}\|_2^2\|\mathbb{E}_0\hat{\Sigma}_x - \Gamma_n\|_2 - \sigma_{0n}^2\|G_n\|_2 - \frac{j\varepsilon_n}{4}.\end{aligned} \tag{8.11}$$

Recalling the definition of $\hat{\Sigma}_x$ and invoking Lemma 1.1 in the supplemental document,

$$\begin{aligned}\|\Lambda_{0n}\|_2^2\|\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2 &= \left\|\frac{\Lambda_{0n}}{\sqrt{c_n}}\right\|_2^2\left\|\frac{1}{c_n}\Lambda_{0n}^{\mathrm{T}}(\Sigma_n - \Sigma_{0n})\Lambda_{0n}\right\|_2 \\ &\geq \left\|\frac{\Lambda_{0n}}{\sqrt{c_n}}\right\|_2^2 s_{\min}\left(\frac{\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}}{c_n}\right)\|\Sigma_n - \Sigma_{0n}\|_2 \\ &\geq \left\|\frac{\Lambda_{0n}}{\sqrt{c_n}}\right\|_2^2 s_{\min}\left(\frac{\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}}{c_n}\right)\frac{j\varepsilon_n}{2}.\end{aligned} \tag{8.12}$$

The last inequality in (8.12) used the triangle inequality to obtain $\|\Sigma_n - \Sigma_{0n}\|_2 \geq \|\Sigma_{1n} - \Sigma_{0n}\|_2 - \|\Sigma_n - \Sigma_{1n}\|_2 \geq j\varepsilon_n - j\varepsilon_n/2 = j\varepsilon_n/2$. By (A3) and

(8.6), both $\|\Lambda_{0n}/\sqrt{c_n}\|_2$ and $s_{\min}(\Lambda_{0n}^{\mathrm{T}}\Lambda_{0n}/c_n)$ can be bounded below by $4/5$. Hence, $\|\Lambda_{0n}\|_2^2\|\hat{\Sigma}_x - \mathbb{E}_{\Sigma_n}\hat{\Sigma}_x\|_2$ is bounded below by $64j\varepsilon_n/250$. Further, based on the calculations following (8.8),

$$\|\Lambda_{0n}\|_2^2\|\mathbb{E}_0\hat{\Sigma}_x - \Gamma_n\|_2 + \sigma_{0n}^2\|G_n\|_2 + \frac{j\varepsilon_n}{4}$$

can be bounded above $63j\varepsilon_n/250$ for n large enough. Substituting in (8.11),

$$\mathbb{P}_{\Sigma_n}[\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 \le j\varepsilon_n/2] < \mathbb{P}_{\Sigma_n}[\|\Lambda_{0n}\|_2^2\|\hat{\Sigma}_x - \mathbb{E}_{\Sigma_n}\hat{\Sigma}_x\|_2 > Cj\varepsilon_n].$$

As in case of the type-I error, using $\|\Lambda_{0n}\|_2^2 \lesssim c_n$, we conclude that

$$\mathbb{E}_{\Sigma_n}(1-\phi_{j,n}) \le \mathbb{P}_{\Sigma_n}[\|\hat{\Sigma}_x - \mathbb{E}_{\Sigma_n}\hat{\Sigma}_x\|_2 > Cj\varepsilon_n/c_n]. \tag{8.13}$$

We are now in a position to invoke (8.5) to bound the right-hand side of (8.13). Using triangle inequality and von Neumann's trace inequality [34],[6]

$$\begin{aligned}\operatorname{tr}(\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x) &\le \operatorname{tr}(\mathbb{E}_0\hat{\Sigma}_x) + |\operatorname{tr}(\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x)| \\ &\le \operatorname{tr}(\mathbb{E}_0\hat{\Sigma}_x) + k_{0n}\|\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2.\end{aligned}$$

Since $\Sigma_n \in B_{j,n}$, $\|\Sigma_n - \Sigma_{0n}\|_2 \le (j+1)\varepsilon_n < 2j\varepsilon_n$, and hence

$$\begin{aligned}\|\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x - \mathbb{E}_0\hat{\Sigma}_x\|_2 &= \|\Lambda_{0n}^{\mathrm{T}}(\Sigma_n - \Sigma_{0n})\Lambda_{0n}\|_2/c_n^2 \\ &\le C\|\Sigma_n - \Sigma_{0n}\|_2/c_n \le 2Cj\varepsilon_n/c_n.\end{aligned}$$

Substituting in the previous display and using $\operatorname{tr}(\mathbb{E}_0\hat{\Sigma}_x) \le Ck_{0n}$, one has $\operatorname{tr}(\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x) \lesssim k_{0n}\max\{1, j\varepsilon_n/c_n\} \le Ck_{0n}(j/\log j)$, with the last inequality using $\varepsilon_n \log j \le c_n$.

Choosing $s = Cn(\log j)^2\varepsilon_n^2/(k_{0n}^2c_n^2)$, one has $s \gtrsim \log k_{0n}$ and $C\operatorname{tr}(\mathbb{E}_{\Sigma_n}\hat{\Sigma}_x) \times \sqrt{(s+\log k_{0n})/n} \le Cj\varepsilon_n/c_n$. By (8.5), the expression in the right-hand side of (8.13) is bounded above by $e^{-s} = \exp\{-Cn(\log j)^2\varepsilon_n^2/(k_{0n}^2c_n^2)\}$. Since the bound is independent of $\Sigma_n \in E_{j,n}$, (8.4) follows. $\square$

9. Proof of the main results.

We now proceed to prove the results stated in Section 5. We prove Theorem 5.3 with the shrinkage prior (PS); the special case of $k_{0n} = O(1)$ in Theorem 5.2 follows immediately. For the point mass prior, we only sketch an argument. We introduce a number of auxiliary Lemmata 9.1, 9.2, 9.3 whose proofs can be found in the supplemental document.

[6] $|\operatorname{tr}(A)| \le k\|A\|_2$ for a $k \times k$ matrix A.

9.1. *Proof of Theorem 5.3.* Set $\varepsilon_n = c_n k_{0n}^{3/2}\sqrt{\frac{s_n \log p_n}{n}}\sqrt{\log n}$ and define $U_n = \{\Sigma_n : \|\Sigma_n - \Sigma_{0n}\|_2 \le M\varepsilon_n\}$. The posterior probability assigned to the complement of U_n is given by

$$\Pi_n(U_n^c \mid \mathbf{y}^{(n)}) = \frac{\int_{U_n^c} \prod_{i=1}^n (f_{\Sigma_n}(y_i)/f_{\Sigma_{0n}(y_i)})\, d\Pi_n(\Sigma_n)}{\int \prod_{i=1}^n (f_{\Sigma_n}(y_i)/f_{\Sigma_{0n}(y_i)})\, d\Pi_n(\Sigma_n)} \equiv \frac{\mathcal{N}_n}{\mathcal{D}_n}, \tag{9.1}$$

where f_{Σ_n} denotes a p_n-dimensional $\mathrm{N}(0, \Sigma_n)$ distribution and $\mathcal{N}_n$ and $\mathcal{D}_n$ denote the numerator and denominator of the fraction in (9.1).

Let $\sigma(y_1, \dots, y_n)$ denote the σ-field generated by $y_1, \dots, y_n$. We first claim that we can lower-bound $\mathcal{D}_n$ on an event $A_n \in \sigma(y_1, \dots, y_n)$ with large probability under $f_{\Sigma_{0n}}$ in Lemma 9.1.

LEMMA 9.1. *Let $\Sigma_{0n} \in \mathcal{C}_{0n}$. Let η_n be a sequence satisfying $\eta_n/s_{\min}(\Sigma_{0n}) \to 0$ and $n\eta_n^2/s_{\min}(\Sigma_{0n})^2 \to \infty$, and define $\varrho_n = 2s_{\max}(\Sigma_{0n})/s_{\min}(\Sigma_{0n})$. Then there exists $A_n \in \sigma(y_1, y_2, \dots, y_n)$ with $\mathbb{P}_{\Sigma_{0n}}(A_n) \to 1$ such that on A_n,*

$$\mathcal{D}_n \ge e^{-Cn\eta_n^2 \log(\varrho_n)/s_{\min}(\Sigma_{0n})^2}\Pi_n(\Sigma_n : \|\Sigma_n - \Sigma_{0n}\|_F < \eta_n).$$

We shall set $\eta_n = \sqrt{s_n k_{0n}/n}$ in all future usage of Lemma 9.1. Based on our prior specification, Σ_n can be parameterized by $(k, \Lambda_n, \sigma_n^2)$ with $k \in \{1, \dots, \infty\}, \Lambda_n \in \mathbb{R}^{p_n \times k}, \sigma_n^2 \in (0, \infty)$, and $\Sigma_n = \Lambda_n\Lambda_n^{\mathrm{T}} + \sigma_n^2 \mathrm{I}_{p_n}$. We use this to bound $\Pi_n(\|\Sigma_n - \Sigma_{0n}\|_F \le \eta_n)$ from below in the following Lemma 9.2.

LEMMA 9.2. *If $\Sigma_{0n} \in \mathcal{C}_{0n}$, the prior Π_n on Σ_n is as in Theorem 5.2, and $\eta_n = \sqrt{s_n k_{0n}/n}$, then*

$$\Pi_n(\|\Sigma_n - \Sigma_{0n}\|_F \le \eta_n) \ge e^{-Cs_n k_{0n} \log n}.$$

We now introduce some notation. Let

$$\begin{aligned} &e_n = s_n k_{0n} \log n, \qquad t_n = Ce_n^2, \qquad \delta_n = \varepsilon_n/(e_n t_n), \\ &\delta_n' = \delta_n/(p_n e_n). \end{aligned} \tag{9.2}$$

Recalling that the true loadings has $s_n k_{0n}$ many nonzero entries, e_n can be thought of as an *effective sparsity* parameter. Also, recall the supp_δ notation from Section 7. Given k, let $\mathrm{supp}_{\delta_n'}(\Lambda_n)$ denote the set $S \subset \{1, \dots, p_n k\}$ corresponding to the entries in $\mathrm{vec}(\Lambda_n)$ larger than δ_n' in absolute magnitude.

Since $\mathbb{P}_{\Sigma_{0n}}(A_n) \to 1$ by Lemma 9.1, it is enough to show

$$\lim_{n\to\infty} \mathbb{E}_0[\Pi_n(U_n^c \mid \mathbf{y}^{(n)}) 1_{A_n}] = 0$$

to prove Theorem 5.2, where $\mathbb{E}_0$ is a shorthand for $\mathbb{E}_{\Sigma_{0n}}$. For some $H > 0$ to be chosen later,

$$\begin{aligned}\mathbb{E}_0[\Pi_n(U_n^c \mid \mathbf{y}^{(n)})1_{A_n}] \le{}& \mathbb{E}_0[\Pi_n(U_n^* \mid \mathbf{y}^{(n)})1_{A_n}] \\ &+ \mathbb{E}_0[\Pi_n(W_n^c \cap V_n \mid \mathbf{y}^{(n)})1_{A_n}] \\ &+ 2\mathbb{E}_0[\Pi_n(V_n^c \mid \mathbf{y}^{(n)})1_{A_n}],\end{aligned} \tag{9.3}$$

where $U_n^* = U_n^c \cap W_n \cap V_n$, with

$$\begin{aligned} W_n &= \{|\mathrm{supp}_{\delta_n'}(\Lambda_n)| \le He_n, \|\Lambda_n\|_1 \le t_n, \sigma^2 \le t_n\}, \\ V_n &= \{k \le Ce_n\}.\end{aligned} \tag{9.4}$$

Thus, U_n^* consists of (strictly speaking, can be identified with the class of) covariance matrices $\Sigma_n = \Lambda_n\Lambda_n^{\mathrm{T}} + \sigma_n^2 \mathrm{I}_{p_n}$ satisfying $\|\Sigma_n - \Sigma_{0n}\|_2 > M\varepsilon_n$, where $\Lambda_n \in \mathbb{R}^{p_n\times k}$ with $k \le Ce_n$, $\|\Lambda_n\|_1 \le t_n$, $|\mathrm{supp}_{\delta_n'}(\Lambda_n)| \le He_n$ and $\sigma_n^2 \le t_n$.

We now show in Lemma 9.3 that the expression in (9.3) goes to zero, so that we can focus on $\Pi_n(U_n^* \mid \mathbf{y}^{(n)})$. This will be crucial in reducing the entropy of the model space.

LEMMA 9.3. *Recall the sequences and sets in (9.2) and (9.4), respectively. There exist constants $H, C > 0$ such that*

$$\lim_{n\to\infty} \mathbb{E}_0[\Pi_n(W_n^c \cap V_n \mid \mathbf{y}^{(n)})1_{A_n}] = 0,$$

$$\lim_{n\to\infty} \mathbb{E}_0[\Pi_n(V_n^c \mid \mathbf{y}^{(n)})1_{A_n}] = 0.$$

For $k \le Ce_n$, a set $S \subset \{1, \dots, p_n k\}$ with $|S| \le He_n$, and $j \ge M$, let $B_{k,S,j,n}$ denote the following subset of U_n^*:

$$\begin{aligned} B_{k,S,j,n} = \{&\Sigma_n = \Lambda_n\Lambda_n^{\mathrm{T}} + \sigma_n^2\mathrm{I}_{p_n} : \Lambda_n \in \mathbb{R}^{p_n\times k}, k \le Ce_n, \|\Lambda_n\|_1 \le t_n, \\ &\sigma_n^2 \le t_n, \mathrm{supp}_{\delta_n'}(\Lambda_n) = S, j\varepsilon_n \le \|\Sigma_n - \Sigma_{0n}\|_2 < (j+1)\varepsilon_n\}.\end{aligned} \tag{9.5}$$

Then, using a standard testing argument (see, e.g., the proof of Proposition 5.1 in [17]),

$$\begin{aligned} &\mathbb{E}_0[\mathbb{P}(U_n^* \mid \mathbf{y}^{(n)})1_{A_n}] \\ &\quad \le \sum_{k\le Ce_n} \sum_{S : |S| \le He_n} \sum_{j\ge M} \Big[\mathbb{E}_0\Phi_{k,S,j,n} \\ &\qquad\qquad + \beta_{k,S,j,n} \sup_{\Sigma_n \in B_{k,S,j,n}} \mathbb{E}_{\Sigma_n}(1 - \Phi_{k,S,j,n})\Big],\end{aligned} \tag{9.6}$$

where $\Phi_{k,S,j,n}$ is a (point vs. composite) test function for

$$H_0 : \Sigma_n = \Sigma_{0n} \quad \text{versus} \quad H_1 : \Sigma_n \in B_{k,S,j,n} \tag{9.7}$$

whose construction is provided below and

$$\beta_{k,S,j,n} := \frac{\Pi_n(B_{k,S,j,n})}{e^{-n\eta_n^2 \log \varrho_n / s_{\min}(\Sigma_{0n})^2} \Pi_n(\|\Sigma_n - \Sigma_{0n}\|_F < \eta_n)} \le e^{Ce_n}. \tag{9.8}$$

To obtain the upper bound on $\beta_{k,S,j,n}$ in the above display, bound $\Pi_n(B_{k,S,j,n})$ above by 1, use the fact that $\log \varrho_n \lesssim \log c_n \le \log n$ by (A1) and (A4) and use Lemma 9.2 to conclude that $\beta_{k,S,j,n} \le e^{Ce_n}$.

To construct the test function $\Phi_{k,S,j,n}$ in (9.6), we cover $B_{k,S,j,n}$ with a union of balls and obtain local tests for Σ_{0n} versus the centers of each of the balls using Theorem 8.1. Since we are inside $W_n \cap V_n$, the number of such balls can be controlled and $\Phi_{k,S,j,n}$ is obtained as the maximum of the local tests.

Let $\Sigma_{n,l}$[7] for $l \in I_{k,S,j,n}$ be a $j\varepsilon_n/2$-net of $B_{k,S,j,n}$ in *operator norm* and for each l, define $E_{k,S,j,n,l} = \{\Sigma_n \in B_{k,S,j,n} : \|\Sigma_n - \Sigma_{n,l}\|_2 \le j\varepsilon_n/2\}$. By definition,

$$B_{k,S,j,n} \subset \bigcup_{l \in I_{k,S,j,n}} E_{k,S,j,n,l}.$$

Clearly, $j\varepsilon_n \le \|\Sigma_{n,l} - \Sigma_{0n}\|_2 < (j+1)\varepsilon_n$ and $\varepsilon_n \ge c_n k_{0n}\sqrt{\log k_{0n}/n}$. For $\Sigma_n \in U_n^*$, $\|\Lambda_n\|_2 \le \sqrt{e_n} t_n$ and $\|\Sigma_n - \Sigma_{0n}\|_2 \le \|\Sigma_n\|_2 + c_n \le e_n t_n^2 + t_n + c_n \lesssim e_n^5$. Hence, $B_{k,S,j,n} \subset U_n^*$ implies $j \lesssim e_n^5/\varepsilon_n$, and hence $\log j \lesssim \log n$, so that $\varepsilon_n \log j \le c_n$ by the second part of (A1). Therefore, the conditions of Theorem 8.1 for the *point versus composite* test $H_0 : \Sigma_n = \Sigma_{0n}$ versus $H_1 : \Sigma_n \in E_{k,S,j,n,l}$ are satisfied. Let $\phi_{k,S,j,n,l}$ denote the corresponding test function from Theorem 8.1 with type-I error $e^{-Cnj^2\varepsilon_n^2/(c_n^2 k_{0n}^2)} = e^{-Cj^2 e_n \log p_n}$ and type-II error $e^{-Cn(\log j)^2\varepsilon_n^2/(c_n^2 k_{0n}^2)} = e^{-C(\log j)^2 e_n \log p_n}$. Letting $\Phi_{k,S,j,n} = \max_{l \in I_{k,S,j,n}} \phi_{k,S,j,n,l}$, we therefore have

$$\mathbb{E}_0(\Phi_{k,S,j,n}) \le |I_{k,S,j,n}| e^{-Cj^2 e_n \log p_n},$$

$$\sup_{\Sigma_n \in B_{j,S,n}} \mathbb{E}_{\Sigma_n}(1 - \Phi_{k,S,j,n}) \le e^{-C(\log j)^2 e_n \log p_n}.$$

To estimate $|I_{k,S,j,n}|$, that is, the covering number of $B_{k,S,j,n}$ in operator norm, we embed $B_{k,S,j,n}$ inside $\tilde{B}_{k,S,j,n}$, whose covering number is easier to calculate:

$$\begin{aligned}\tilde{B}_{k,S,j,n} := \{\Sigma_n = \Lambda_n\Lambda_n^{\mathrm{T}} + \sigma_n^2 \mathrm{I}_{p_n} : \Lambda_n \in B_{k,S,j,n}^{(\Lambda)}, \sigma_n^2 \le t_n,\\ j\varepsilon_n \le \|\Sigma_n - \Sigma_{0n}\|_2 < (j+1)\varepsilon_n\},\end{aligned}$$

where $B_{k,S,j,n}^{(\Lambda)} = \{\Lambda_n \in \mathbb{R}^{p_n \times k} : k \le Ce_n, \mathrm{supp}_{\delta_n'}(\Lambda_n) = S, \|\Lambda_n\|_F \le e_n t_n\}$. The containment $B_{k,S,j,n} \subset \tilde{B}_{k,S,j,n}$ follows since $\|\Lambda_n\|_F \le \sqrt{k}\|\Lambda_n\|_2 \le k\|\Lambda_n\|_1 \le e_n t_n$.

[7]We suppress the dependence on k, S and j.

We now proceed to explicitly construct a $j\varepsilon_n/2$-net for $\tilde{B}_{k,S,j,n}$. Let $\xi_n = j\varepsilon_n/(8e_n t_n)$. For notational convenience, we use $P_S(\theta)$ below to denote θ_S defined in Section 2. Let $\{\Lambda_l\}_{l=1}^L$ be a ξ_n-net of $B_{k,S,j,n}^{(\Lambda)}$. Also, let $\{\sigma_r^2\}_{r=1}^R$ be a $j\varepsilon_n/4$-net of $[0, t_n]$. We show below that $\{\Lambda_l\Lambda_l^{\mathrm{T}} + \sigma_r^2\}_{l,r}$ form a $j\varepsilon_n/2$-net of $\tilde{B}_{k,S,j,n}$ in operator norm.

Let $\tilde{\Sigma} = \tilde{\Lambda}\tilde{\Lambda}^{\mathrm{T}} + \tilde{\sigma}^2\mathrm{I}$ be in $\tilde{B}_{k,S,j,n}$. Find Λ_l and σ_r^2 from the respective nets so that $\|\Lambda_l - \tilde{\Lambda}\|_F \le \xi_n$ and $|\sigma_r^2 - \tilde{\sigma}^2| \le j\varepsilon_n/4$. Let $\Sigma = \Lambda_l\Lambda_l^{\mathrm{T}} + \sigma_r^2$. Then

$$\|\Sigma - \tilde{\Sigma}\|_2 \le j\varepsilon_n/4 + \|\Lambda_l\Lambda_l^{\mathrm{T}} - \tilde{\Lambda}\tilde{\Lambda}^{\mathrm{T}}\| \le j\varepsilon_n/4 + [\|\Lambda_l\|_2 + \|\tilde{\Lambda}\|_2]\xi_n \le j\varepsilon_n/2.$$

We have thus proved our claim, and hence $|I_{k,S,j,n}| \le L \times R$. Note the use of the control on $\|\Lambda\|_2$ over $B_{k,S,j,n}$ in the above display.

Clearly, R can be chosen to smaller than $t_n/(2j\varepsilon_n)$. With $s = |S|$, let $\{\theta_l\}_{l=1}^L$ be a $\xi_n/2$-net of the Euclidean sphere in $\mathbb{R}^s$ of radius $e_n t_n$. By Lemma 5.2 of [40], the cardinality of such a net L can be chosen to be smaller than $(1 + e_n t_n/\xi_n)^s$. We now exhibit a ξ_n-net $\{\Lambda_l\}_{l=1}^L$ for $B_{k,S,j,n}^{(\Lambda)}$ in Frobenius norm as follows. Set $P_S(\Lambda_l) = \theta_l$ and $P_{S^c}(\Lambda_l) = \mathbf{0}$. Let $\Lambda \in B_{k,S,j,n}^{(\Lambda)}$ and $\theta = P_S(\Lambda)$. There exists θ_l such that $\|\theta - \theta_l\|_2 \le \xi_n/2$. Also, since $\mathrm{supp}_{\delta_n'}(\Lambda) = S$, $\|P_{S^c}(\Lambda)\|_2 \le \delta_n$. By choosing j larger than some constant J, we can make $\xi_n \ge 2\delta_n$. Hence, $\|\Lambda_l - \Lambda\|_F \le \xi_n$. Thus, $L \times R$ can be bounded above by $e^{Cs\log(e_n t_n)} \le e^{Cs\log p_n}$, and hence

$$\mathbb{E}_0(\Phi_{k,S,j,n}) \le e^{Cs\log p_n} e^{-C_1 j^2 e_n \log p_n}, \tag{9.9}$$

$$\sup_{\Sigma_n \in B_{k,S,j,n}} \mathbb{E}_{\Sigma_n}(1 - \Phi_{k,S,j,n}) \le e^{-C_2(\log j)^2 e_n \log p_n}. \tag{9.10}$$

Substitute the bounds obtained in (9.8), (9.9) and (9.10) in (9.6). Observing that all the bounds are free of k, we can bound the expression in (9.6) by

$$(Ce_n)\sum_{s=0}^{He_n}\binom{p_n}{s}\left[\sum_{j\ge M} e^{Cs\log p_n} e^{-C_1 j^2 e_n \log p_n} + e^{Ce_n} e^{-C_2(\log j)^2 e_n \log p_n}\right]. \tag{9.11}$$

The first term in the inner sum over j can be bounded above by $e^{Ce_n\log p_n} \times e^{-C_3 M^2 e_n \log p_n}$, while the second one by $e^{Ce_n} e^{-C_4(\log M)^2 e_n \log p_n}$. Noting that $(Ce_n) \times (He_n + 1)\max_{\{0\le l\le He_n\}}\binom{p_n}{l} \le \exp\{Ce_n \log p_n\}$, (9.11) goes to 0 as $n \to \infty$ for a large enough constant $M > 0$. This completes the proof of Theorem 5.3 with the shrinkage prior (PS).

The proof for the point mass priors (PL1) follows similarly. Since the point mass mixture priors allow exact zeros in the loadings, we can condition on $\mathrm{supp}(\Lambda_n) = S$. By properties of point mass mixture priors shown in [17], analogues of Lemmata 9.2 and 9.3 can be obtained to complete the theorem.

APPENDIX

PROOF OF LEMMA 5.5. Observe that if $\|\theta_j - \theta_{j'}\|_H = 2(s-r)$, then $\langle b_j, b_{j'}\rangle = r$. For $j \neq j'$, $\Sigma_{(j)} - \Sigma_{(j')} = \gamma(b_j b_j^{\mathrm{T}} - b_{j'} b_{j'}^{\mathrm{T}})$. The nonzero eigenvalues of the matrix $B = (b_j b_j^{\mathrm{T}} - b_{j'} b_{j'}^{\mathrm{T}})$ are $\{\sqrt{s^2-r^2}, -\sqrt{s^2-r^2}\}$, since $\mathrm{rk}(B) = 2$, $\mathrm{tr}(B) = 0$ and $\mathrm{tr}(B^2) = 2(s^2-r^2)$.

Since $\theta_j \in \mathcal{M}$ for all j, by symmetry, $\det(\Sigma_{(j)}) = \det(\Sigma_{(j')})$ for all $j \neq j'$. Hence, $\mathrm{KL}(\mathbb{P}^{(j)}, \mathbb{P}^{(j')}) = (n/2)\{\mathrm{tr}(\Sigma_{(j)}^{-1}\Sigma_{(j')}) - p\}$. Write $\Sigma_{(j)} = \beta(A + t b_j b_j^{\mathrm{T}})$, where A is a diagonal matrix with the first $(p-1)$ diagonal entries equaling one and the pth entry being $(1+\kappa/\beta)$. An application of the Woodbury matrix inversion formula produces

$$(A + t b_j b_j^{\mathrm{T}})^{-1} = A^{-1} - \frac{t}{1+ts} b_j b_j^{\mathrm{T}},$$

so that

$$\Sigma_{(j)}^{-1}\Sigma_{(j')} = \mathrm{I}_p - \frac{t}{1+ts} b_j b_j^{\mathrm{T}} + t b_{j'} b_{j'}^{\mathrm{T}} - \frac{t^2 r}{1+ts} b_j b_{j'}^{\mathrm{T}}.$$

The proof is completed by observing that $\mathrm{tr}(b_j b_j^{\mathrm{T}}) = s$ and $\mathrm{tr}(b_j b_{j'}^{\mathrm{T}}) = r$. $\square$

PROOF OF LEMMA 5.6. Let $\tau \in \mathcal{M}$ with $\mathrm{supp}(\tau) = S$. We show that for any $x \in \mathcal{M}$, $\|x-\tau\|_H = 2\sum_{j\notin S} 1(x_j = 1)$. To that end, we have $\|x-\tau\|_H = \sum_{j\in S} 1(x_j = 0) + \sum_{j\notin S} 1(x_j = 1) = s + a - b$, where $a = \sum_{j\notin S} 1(x_j = 1)$, $b = \sum_{j\in S} 1(x_j = 1)$. Since $x \in \mathcal{M}$, we also have $a + b = s$, which implies $\|x-\tau\|_H = 2a$.

Let k denote the integer part of $s/6$. Let $\mathcal{M}_0$ be a maximal set of points in $\mathcal{M}$, with each pair at least $2(k+1)$ apart in Hamming distance. Note here that $2(k+1) > s/3$. Since $\mathcal{M}_0$ is maximal and $d(x,y)$ is even for any $x, y \in \mathcal{M}$ by the above calculation, it follows that $\mathcal{M} \subset \bigcup_{\tau\in\mathcal{M}_0} B(\tau; 2k)$, where

$$B(\tau; 2k) = \{x \in \mathcal{M} : \|x-\tau\|_H \le 2k\}.$$

By symmetry, $B(\tau; 2k)$ is independent of τ, so that $|\mathcal{M}| \le |\mathcal{M}_0||B(\tau; 2k)|$ for any $\tau \in \mathcal{M}_0$. It is easy to see that

$$|B(\tau; 2k)| = \sum_{j=0}^{k} |A_j| = \sum_{j=0}^{k} \binom{s}{j}\binom{q-s}{j},$$

where $A_j = \{x \in \mathcal{M} : \|x-\tau\| = 2j\}, 0 \le j \le k$. Since $k \le s/2$, the expression in the above display can be bounded above by $k\binom{s}{k}\binom{p-1}{k}$. One thus has $|\mathcal{M}| = \binom{p-1}{s} \le mk\binom{s}{k}\binom{p-1}{k}$. Using $(n/r)^r \le \binom{n}{r} \le (ne/r)^r$ for $0 \le r \le n/2$, we obtain $m \geq \exp(Cs\log p)$ for some constant $C > 0$. Also, clearly $m \leq |\mathcal{M}| \leq \exp(C_1 s \log p)$. $\square$

PROOF OF LEMMA 7.1. Let $\delta = \varepsilon/p$. To lower-bound $\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon)$, we first obtain a lower bound conditioned on the hyper parameters τ and γ:

$$\begin{aligned}
&\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon \mid \tau, \gamma) \\
&\quad \geq \mathbb{P}(|\theta_j| \leq \delta \ \forall j \in S_0^c \mid \tau, \gamma)\mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < \varepsilon/2 \mid \tau, \gamma) \qquad \text{(A.1)}\\
&\quad = \Bigg[\prod_{j \in S_0^c} (1 - e^{-\delta/\psi_j})\Bigg] \times \mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < \varepsilon/2 \mid \tau, \gamma).
\end{aligned}$$

Let $\tilde{\gamma} = (\gamma_1, \ldots, \gamma_{p-1})^{\mathrm{T}}$ and $\gamma_p = 1 - \sum_{j=1}^{p-1} \gamma_j$. We now have to integrate out τ and $\tilde{\gamma}$ in (A.1). By a relabeling of indices, we can always make sure that the pth index lies in S_0. Let $S_1 = S_0 \setminus \{p\}$ so that $S_0^c \cup S_1 = \{1, \ldots, p-1\}$. Fix numbers $a, b \in (0,1)$ with $b = 4a$. Observe that if $\tau \in [2s, 4s]$, $\gamma_j \tau \leq \frac{\delta}{\log(p/s)}$ $\forall j \in S_0^c$ and $\gamma_j\tau \in [a,b]$ $\forall j \in S_1$, then for $\varepsilon < b/2$,

$$\sum_{j=1}^{p-1} \gamma_j = \sum_{j \in S_0^c} \gamma_j + \sum_{j \in S_1} \gamma_j \leq \varepsilon + \frac{(s-1)b}{2s} \leq b < 1. \qquad \text{(A.2)}$$

Define $\mathcal{B} \subset \Delta_0^{p-1} \times \mathbb{R}^+$ such that

$$\mathcal{B} = \Bigg\{(\gamma, \tau) : 0 \leq \gamma_j\tau \leq \frac{\delta}{\log(p/s)} \ \forall j \in S_0^c; \gamma_j\tau \in [a,b] \ \forall j \in S_1, \quad \tau \in [2s, 4s]\Bigg\}. \qquad \text{(A.3)}$$

Clearly, $\mathcal{B}$ is a measurable subset of $\Delta_0^{p-1} \times \mathbb{R}^+$. For a fixed τ in the interval $[2s, 4s]$, the section $\mathcal{A}_\tau \subset \Delta_0^{p-1}$ is given by

$$\mathcal{A}_\tau = \Bigg\{0 \leq \gamma_j \leq \frac{\delta}{\log(p/s)\tau} \ \forall j \in S_0^c; \gamma_j \in \bigg[\frac{a}{\tau}, \frac{b}{\tau}\bigg] \ \forall j \in S_1\Bigg\}. \qquad \text{(A.4)}$$

Thus,

$$\begin{aligned}
\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon) &= \int_{(\tau, \tilde{\gamma}) \in \mathbb{R}^+ \times \Delta_0^{p-1}} \mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon \mid \tau, \tilde{\gamma}) f_\gamma(d\tilde{\gamma}) f_\tau(d\tau) \qquad \text{(A.5)}\\
&\geq \int_{(\tau, \gamma) \in \mathcal{B}} \mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon \mid \tau, \gamma) f_\gamma(d\tilde{\gamma}) f_\tau(d\tau).
\end{aligned}$$

We now substitute the lower bound for $\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon \mid \tau, \gamma)$ from (A.1) in (A.5) and lower-bound the two terms on the right-hand side of (A.1) individually.

For the first term, observe that for $(\tau,\gamma)\in\mathcal{B}$, $\prod_{j\in S_0^c}(1-e^{-\delta/\psi_j})\geq(1-s/p)^{p-s}$.

To tackle the second term, we make use of Lemma 7.2. By definition, $\psi_j\in[a,b]$ for all $j\in S_1$ whenever $(\tau,\gamma)\in\mathcal{B}$. Further, along the lines of (A.2), $\sum_{j=1}^{p-1}\gamma_j\in[a/8,b]$, and hence $\gamma_p\in[1-b,1-a/8]$ on $\mathcal{B}$. Hence, $\psi_p\in[2s(1-b),4s(1-a/8)]$. Since a,b are constants, by a slight abuse of notation, we shall assume $\psi_j\in[a,b]$ for all $j\in S_1$ and $\psi_p\in[2sa,4sb]$ on $\mathcal{B}$. It thus follows from Lemma 7.2 that

$$\begin{aligned}&\mathbb{P}(\|\Pi_{S_0}(\theta)-\Pi_{S_0}(\theta_0)\|_2<\varepsilon/2\mid\tau,\gamma)\\&\quad\geq\exp\biggl\{-\frac{C_1}{a^2}\sum_{j\in S_0}|\theta_{0j}|^2-C_2s-s|\log\{\varepsilon/(2b\sqrt{s})\}|\biggr\}.\end{aligned}$$

We conclude that for $(\tau,\gamma)\in\mathcal{B}$, the integrand in (A.5) can be bounded below as follows:

$$\begin{aligned}&\mathbb{P}(\|\theta-\theta_0\|_2<\varepsilon\mid\tau,\gamma)\\&\quad\geq e^{-Cs}\exp\biggl\{-\frac{C_1}{a^2}\sum_{j\in S_0}|\theta_{0j}|^2-C_2s-s|\log\{\varepsilon/(2b\sqrt{s})\}|\biggr\},\end{aligned}\tag{A.6}$$

where the last inequality uses $(1-x)^{1/x}\geq1/(2e)$ for $0\leq x\leq1/2$ and $C=\log(2e)$. It thus remains to obtain a lower bound to

$$\mathbb{P}(\mathcal{B})=\int_{(\tau,\gamma)\in\mathcal{B}}f_\gamma(d\tilde{\gamma})f_\tau(d\tau)=\int_{\tau=2s}^{4s}\mathbb{P}(\mathcal{A}_\tau\mid\tau)f_\tau(d\tau).\tag{A.7}$$

Now, since $\gamma\sim\mathrm{Dir}(\alpha/p,\dots,\alpha/p)$, recalling the definition of $\mathcal{A}_\tau$ from (A.4) and using (A.2),

$$\begin{aligned}&\mathbb{P}(A_\tau\mid\tau)\\&\quad=\frac{\Gamma(\alpha)}{\Gamma(\alpha/p)^p}\int_{\tilde{\gamma}\in\mathcal{A}_\tau}\Biggl[\prod_{j=1}^{p-1}\gamma_j^{\alpha/p-1}\Biggr]\Biggl(1-\sum_{j=1}^{p-1}\gamma_j\Biggr)^{\alpha/p-1}d\gamma_1\cdots d\gamma_{p-1}\\&\quad\geq C_p(1-b)^{\alpha/p-1}\int_{\tilde{\gamma}\in\mathcal{A}_\tau}\Biggl[\prod_{j\in S_1}\gamma_j^{\alpha/p-1}\Biggr]\times\Biggl[\prod_{j\in S_0^c}\gamma_j^{\alpha/p-1}\Biggr]d\gamma_1\cdots d\gamma_{p-1}\\&\quad\geq C_p(1-b)^{\alpha/p-1}\biggl\{\frac{\delta}{\log(p/s)}\biggr\}^{\alpha(p-s)/p}\biggl\{\biggl(\frac{b}{\tau}\biggr)^{\alpha/p}-\biggl(\frac{a}{\tau}\biggr)^{\alpha/p}\biggr\}^{s-1},\end{aligned}\tag{A.8}$$

where

$$C_p=\frac{\Gamma(\alpha)}{\Gamma(\alpha/p)^p}\biggl(\frac{p}{\alpha}\biggr)^{p-1}$$

$$
\begin{aligned}
&= \exp\{\log\Gamma(\alpha) + (p-1)\log(p/\alpha) - p\log\Gamma(\alpha/p)\} \\
&\geq \exp\{\log\Gamma(\alpha) - \log\Gamma(\alpha/p)\} \\
&\geq \exp\{\log\Gamma(\alpha) - \log(p/\alpha)\}
\end{aligned}
\tag{A.9}
$$

with the last two inequalities using $\Gamma(x) \leq 1/x$ for all $x \in (0,1)$. Moreover, since $b \geq 4a$, we have for $\tau \in [2s, 4s]$,

$$
\begin{aligned}
\left\{\left(\frac{b}{\tau}\right)^{\alpha/p} - \left(\frac{a}{\tau}\right)^{\alpha/p}\right\}^{s-1} &\geq \left\{\left(\frac{b}{4s}\right)^{\alpha/p} - \left(\frac{a}{2s}\right)^{\alpha/p}\right\}^{s-1} \\
&\geq \left(\frac{b}{4s}\right)^{(s-1)\alpha/p}\left[1 - \exp\left\{-\frac{\alpha}{p}\log(2b/a)\right\}\right].
\end{aligned}
\tag{A.10}
$$

Equations (A.9) and (A.10), in conjunction with the fact that $1 - e^{-x} \geq x/2$ for $x \in (0,1)$ implies that the expression in (A.8), and thus $\mathbb{P}(\mathcal{A}_\tau \mid \tau)$ in (A.7), is bounded below by

$$
\begin{aligned}
&\mathbb{P}(\mathcal{A}_\tau \mid \tau) \\
&\quad \geq C\exp\left\{\frac{\alpha(p-s)}{p}\log\frac{\delta}{\log(p/s)} - \log\frac{p}{\alpha} - \frac{1}{\log(b/2a)}\log\frac{p}{\alpha}\right\}
\end{aligned}
\tag{A.11}
$$

for some constant $C > 0$. Finally, (A.6) and (A.11) substituted into (A.5) gives us

$$
\mathbb{P}(\|\theta - \theta_0\|_2 < \varepsilon) \geq \mathbb{P}[\tau \in (2s, 4s)]e^{-C\max\{\|\theta_0\|_2^2, s\log(s/\varepsilon), \log p\}}.
$$

The proof of Lemma 7.1 is completed upon observing that $\mathbb{P}[\tau \in (2s, 4s)] \geq e^{-Cs}$. $\square$

Proof of Lemma 7.3. Without loss of generality, we provide the proof for $\alpha = 1$. Lemma IV.3 of [43] implies that under (PS),

$$
\theta_j \mid \psi_j \overset{\text{ind.}}{\sim} \mathrm{DE}(\psi_j), \qquad \psi_j \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(1/p, 1/2).
\tag{1.12}
$$

By (1.12), θ_j's are independent and identically distributed, so that $|\mathrm{supp}_\delta(\theta)| \sim \mathrm{Binomial}(p, \zeta)$, with $\zeta := \mathbb{P}(|\theta_1| > \delta)$. We first show that $\zeta \lesssim \log p/p$ for $\delta = \varepsilon/p$. Observe that

$$
\begin{aligned}
&\mathbb{P}(|\theta_1| > \delta) \\
&\quad = \frac{(1/2)^{1/p}}{\Gamma(1/p)}\int_0^\infty e^{-\delta/x}x^{1/p-1}e^{-x/2}\,dx \\
&\quad = \frac{(1/2)^{1/p}}{\Gamma(1/p)}\left\{\int_0^{4\delta} e^{-\delta/x}x^{1/p-1}e^{-x/2}\,dx + \int_{4\delta}^\infty e^{-\delta/x}x^{1/p-1}e^{-x/2}\,dx\right\} \\
&\quad \leq \frac{(1/2)^{1/p}}{\Gamma(1/p)}\left\{C + \int_{4\delta}^\infty \frac{e^{-x/2}}{x}\,dx\right\} \leq \frac{(1/2)^{1/p}}{\Gamma(1/p)}\left\{C + \int_{2\delta}^\infty \frac{e^{-t}}{t}\,dt\right\}.
\end{aligned}
\tag{1.13}
$$

Using a bound for the incomplete gamma function from Theorem 2 of [1],

$$\int_{2\delta}^{\infty} \frac{e^{-t}}{t}\,dt \le -\log(1-e^{-2\delta}) \le -\log(\delta), \tag{1.14}$$

for δ small. Since $\Gamma(1/p) \ge p/2$ for large p, and $C + \log(1/\delta) \le 2\log(1/\delta)$ for p large, we have $\mathbb{P}(|\theta_1| > \delta) \le \log(1/\delta)/p \lesssim \log p/p$; the last inequality follows since $\varepsilon > 1/p^B$ implies $\delta \ge 1/p^{B+1}$.

A version of Chernoff's inequality for the binomial distribution [27] states that for $B \sim \mathrm{Binomial}(p,\zeta)$ and $\zeta \le a < 1$,

$$\mathbb{P}(B > ap) \le \left\{\left(\frac{\zeta}{a}\right)^a e^{a-\zeta}\right\}^p. \tag{1.15}$$

In (1.15), set $a = As/p$. Since $s \gtrsim \log p$, we can ensure $\zeta \le a$ by choosing A larger than some constant. Hence, by (1.15), $\mathbb{P}(|\operatorname{supp}_\delta(\theta)| > As) \le (e\zeta/a)^{As} \le e^{-A\log(A/eC)s}$. $\square$

Proof of Lemma 7.4. Recall $\theta_j \mid \gamma, \tau \sim \mathrm{DE}(\gamma_j\tau)$ for $1 \le j \le p$. Let $X_j = \theta_j/(\gamma_j\tau)$, so that $X_j \mid \gamma,\tau \sim \mathrm{DE}(1)$ independently. Let $\psi_j = \gamma_j\tau$ and fix $t > 1$. Using a Bernstein-type tail inequality for subexponential random variables (Proposition 5.16 of [40]),

$$\begin{aligned}\mathbb{P}\left(\sum_{j=1}^p |\theta_j| > t \,\middle|\, \gamma,\tau\right) &= \mathbb{P}\left(\sum_{j=1}^p |\psi_j X_j| > t \,\middle|\, \gamma,\tau\right)\\ &\le \exp\left\{-C\min\left(\frac{t^2}{\|\psi\|_2^2}, \frac{t}{\|\psi\|_\infty}\right)\right\}\\ &\le \max\{e^{-Ct^2/\tau^2}, e^{-Ct/\tau}\}.\end{aligned}$$

The last inequality in the above display uses $\|\gamma\|_2^2 \le \|\gamma\|_1 = 1$ and $e^{-c/x}$ is increasing in x. Fix $t \ge 1$. Since $\tau \sim \mathrm{Exp}(1/2)$, $\mathbb{P}(\tau > \sqrt{t}) \le e^{-C\sqrt{t}}$. Also, $\max_{0\le x\le\sqrt{t}} \max\{e^{-Ct^2/x^2}, e^{-Ct/x}\} \le e^{-C\sqrt{t}}$. The result follows by noting that

$$\begin{aligned}\mathbb{P}(\|\theta\|_1 > t) &\le \int_{x=0}^{\sqrt{t}} \max\{e^{-Ct^2/x^2}, e^{-Ct/x}\} f_\tau(x)\,dx + \mathbb{P}(\tau > \sqrt{t})\\ &\le 2e^{-C\sqrt{t}}.\end{aligned}$$

$\square$

Acknowledgements. The authors would like to thank two anonymous referees, an Associate Editor and the Editor for their thoughtful comments on previous versions of the paper which has helped improve our exposition. We also thank Steven Finch for careful proofreading of an initial draft of the paper.

REFERENCES

[1] Alzer, H. (1997). On some inequalities for the incomplete gamma function. *Math. Comp.* **66** 771–778. MR1397438

[2] Armagan, A., Dunson, D. and Lee, J. (2011). Generalized double Pareto shrinkage. Available at arXiv:1104.0861.

[3] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

[4] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

[5] Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. Oxford Univ. Press, New York.

[6] Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31** 536–559. MR1983541

[7] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[8] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[9] Birgé, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282. MR0764150

[10] Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.* **39** 2557–2584. MR2906878

[11] Bunea, F. and Xiao, L. (2012). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. Available at arXiv:1212.5321.

[12] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

[13] Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[14] Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

[15] Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722

[16] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[17] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

[18] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[19] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

[20] Fan, J., Liao, Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410

[21] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 603–680. MR3091653

[22] Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** 315–331. MR1681701

[23] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68. MR1790613

[24] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[25] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[26] Giné, E. and Nickl, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395

[27] Hagerup, T. and Rüb, C. (1990). A guided tour of Chernoff bounds. *Inform. Process. Lett.* **33** 305–308. MR1045520

[28] Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *J. Amer. Statist. Assoc.* **106** 1383–1393. MR2896843

[29] Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Ann. Statist.* **35** 1487–1511. MR2351094

[30] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[31] Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663

[32] Le Cam, L. (1986). *Asymptotic methods in statistical decision theory.* Springer, New York.

[33] Lucas, J. E., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R. and West, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics* (K. A. Do, P. Müller and M. Vannucci, eds.) 155–176. Cambridge University Press, Cambridge.

[34] Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatsh. Math.* **79** 303–306. MR0371930

[35] Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

[36] Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford Univ. Press, New York.

[37] Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105

[38] Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

[39] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[40] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Y. C. Eldar and G. Kutyniok, eds.) 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[41] West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics, 7 (Tenerife, 2002)* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 733–742. Oxford Univ. Press, New York. MR2003537

[42] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 423–435. Springer, New York. MR1462963

[43] Zhou, M. and Carin, L. (2012). Negative binomial process count and mixture modeling. Preprint. Available at arXiv:1209.3442.

D. Pati
Department of Statistics
Florida State University
Tallahassee, Florida
USA
E-mail: debdeep@stat.fsu.edu

A. Bhattacharya
Department of Statistics
Texas A&M University
College Station, Texas
USA
E-mail: anirbanb@stat.tamu.edu

N. S. Pillai
Department of Statistics
Harvard University
Cambridge, Massachusetts
USA
E-mail: pillai@fas.harvard.edu

D. Dunson
Department of Statistical Science
Duke University
Durham, North Carolina
USA
E-mail: dunson@stat.duke.edu